SAMMENHENGEN MELLOM LENGDEN AV UTDANNELSE OG DØDELIGHET I NORGE MED FOKUS PÅ SØSKENAVHENGIGHET

av

Narin Khalaf

MASTEROPPGAVE

for graden

Master i Modellering og dataanalyse

Det matematisk- naturvitenskapelige fakultet
Universitetet i Oslo

Mai 2013

Til mine kjære foreldre

Forord

Denne masteroppgaven er blitt skrevet i perioden våren 2012 til våren 2013 og er avslutningen på mastergraden i Modellering og dataanalyse ved Universitetet i Oslo. I løpet av arbeidet med denne oppgaven har jeg hatt noen utfordringer, men samtidig har det også vært både lærerrikt og spennende.

Det er mange personer som har støttet meg underveis. Først og fremst ønsker jeg å takke min veileder, Sven Ove Samuelsen, for å ha gitt meg en interessant oppgave. Han har vært tålmodig og tilgjengelig når jeg har hatt spørsmål. Jeg er takknemlig for alle tilbakemeldingene jeg har fått. I tillegg ønsker jeg å takke Folkehelseinstituttet (FHI) for overlevering av datasett og møter med Øyvind Næss ved FHI angående spørsmål om datasettet.

Jeg ønsker også å takke mine medstudenter på universitetet, venner, kolleger i Mercer og spesielt Ingrid Maudal og Asrunn Aure for å ha lest oppgaven og gitt meg verdifulle tilbakemeldinger. Min sjef Ingrid har vært svært forståelsesfull og gitt meg en fleksibel arbeidstid.

Sist, men ikke minst vil jeg takke mine foreldre og søsken. Mamma og pappa har hele tiden hatt stor tro på meg og støtten deres har betydd uendelig mye. Min tvillingsøster, Nawroz, har vært en stor støtte gjennom hele studietiden.

Oslo, mai 2013
Narin Khalaf

Innhold

Kapittel 1

Innledning

1.1 Bakgrunn

Vi tar utgangspunkt i artikkelen *Education and adult cause specific mortality-examining the impact of family factors shared by 871 367 Norwegian siblings*[1]. I denne studien ble sammenhengen mellom utdannelse og dødelighet i Norge undersøkt, og det var totalt 871 367 personer som deltok. Man har sett på sammenhengen mellom lengden av utdannelse og årsaksspesifikk dødelighet innen søskenflokker og mellom individene. Hovedresultatene fra denne artikkelen er kort oppsummert i dette kapittelet.

1.2 Problemstilling

I denne oppgaven skal vi undersøke betydningen av familieavhengighet innen søskenflokker. For å gjøre dette skal vi skal blant annet bruke Cox regresjonsmodell med og uten stratifisering på søskenflokk. Vi skal også bruke frailty-modeller, som tar hensyn til avhengigheten blant søsken på en annen måte. Videre i oppgaven skal vi anvende bootstrapping for å se om det er systematiske forskjeller mellom de ulike metodene.

1.2.1 Resultater fra artikkelen

De 871 367 personene var fordelt på 337 627 søskenflokker, så grupper på én person ble ikke tatt med i studien. Personer uten informasjon om utdannelse eller der foreldrene var ukjente ble ekskludert fra datasettet.

For å undersøke om søskenavhengigheten har betydning, ble det brukt to metoder med separate analyser for kvinner og menn. De to metodene var Cox regresjon med og uten stratifisering på søskenflokk. I begge metodene tar man hensyn til at det er familieavhengighet innen søskenflokker, men dette gjøres på ulike måter. Siden helsøsken lever i samme miljø og har samme utgangspunkt, er det interessant å undersøke om familieavhengigheten er viktig for dødelighetsrisiko. I den metoden der det ble tilpasset en Cox-modell uten

stratifisering ble *sandwich-estimatoren* brukt. Denne varians-estimatoren tar hensyn til avhengighet mellom søsken. Metoden gir de samme hazard ratioene som når en tilpasser en Cox-modell for uavhengige data, men varians-estimatoren og dermed også konfidensintervaller for hazard ratioene blir ulike. I den andre metoden, stratifisert Cox-modell, ble det stratifisert på søskenflokk. Hver gruppe får da en *baseline-hazard*, i motsetning til den førstnevnte metoden, der baseline er felles for alle personene. Begge metodene er beskrevet mer detaljert i Kapittel 3.

I denne analysen viste resultatene at det var en markant forskjell i risiko for de personene med høy og lav utdannelse for alle dødsårsakene. Dette gjaldt for begge kjønn. For menn gjaldt dette spesielt for dødsårsakene lungekreft og alkoholrelaterte dødsårsaker, mens for kvinner var det CVD og alkoholrelaterte årsaker. I analysene med stratifisert Cox-modell var denne effekten noe svakere. For eksempel ble hazard ratioen 6.29 for menn med 7-9 års utdannelse sammenlignet med menn som hadde universitetsutdannelse for dødsårsaken lungekreft, som tilsvarer en dødsrisiko som er omtrent seks ganger høyere for denne gruppen.

1.2.2 Videre

I denne oppgaven skal vi sammenligne flere metoder, og skal blant annet se på de samme metodene som er brukt i artikkelen. Innledningsvis skal vi se på Cox regresjon, både stratifisert og ustratifisert. Vi skal også anvende marginale modeller. En modell som ikke er brukt i artikkelen er *frailty*-modellen. Frailty-variabelen er en tilfeldig variabel som ikke kan observeres og man antar at alle personene i samme gruppe har samme frailty. Denne modellen forklares nærmere i Kapittel 3. Selv om grupper på en person ikke vil bidra i den stratifiserte analysen, blir de likevel ikke ekskludert fra datasettet i denne oppgaven. Disse personene vil defineres som "enebarn"[1]. Ved å fjerne "enebarn" blir datasettet redusert, og estimatene blir mer sikre jo større datasettet er. For å se om resultatene endres ved at man utelukker "enebarn", skal vi også tilpasse en Cox regresjonsmodell for søsken, slik det er gjort i artikkelen. Videre skal vi i tillegg se på egne analyser for tvillinger og "enebarn". Dette er ikke gjort i artikkelen. For tvillinger er det interessant å undersøke om søskenavhengigheten har større betydning enn blant søsken generelt.

Oppgaven er delt inn slik: I Kapittel 2 vises en oversikt over datasettet. Kapittel 3 er et sammendrag av de metodene og modellene som er brukt i denne oppgaven, og fra fra og med Kapittel 4 til Kapittel 6 vises resultatene fra analysene. Først er det litt om de innledende analysene med Cox regresjonsmodell med og uten stratifisering på søskenflokk for totaldødeligheten, samt alle dødsårsakene. Deretter presenterer vi resultatene fra frailty-analysene der vi bruker gamma og log-normal frailty. Videre sammenlignes de ulike metodene for totaldødelighet og for noen utvalgte dødsårsaker. For å teste om resultatene som fremkommer er tilfeldige eller ikke, er det anvendt boostrapping. Disse resultatene er vist i Kapittel 5. I Kapittel 6 er de samme metodene som i Kapittel 4 og Kapittel 5 benyttet for tvillinger. Det antas at leseren er innforstått med generelle ord og uttrykk innen statistikk.

[1] Definisjonen beskrives nærmere i Kapittel 2.

Kapittel 2

Datasett

I dette kapittelet gis det en beskrivelse av datasettet. Datasettet som brukes i denne oppgaven er en modifisert versjon av det datasettet som ble brukt i studien[1]. Disse forskjellene forklares i neste seksjon. Først beskrives oppfølgingsperioden og kovariatene. Deretter er det flere tabeller med oversikter over antall personer som døde i løpet av den tiden studien varte. Siste del av kapittelet tar for seg hvor mange tvillinger og "enebarn" som var med i datasettet. Grunnen til det er at vi har sett på egne analyser for disse gruppene.

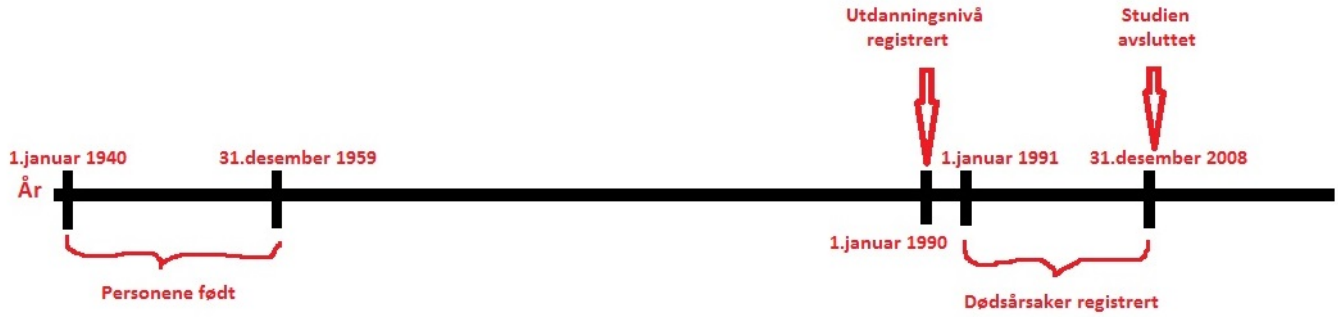
2.1 Beskrivelse av studien

Figur 2.1 viser en illustrasjon av start- og sluttidspunkt for studien. Personene i denne studien var født mellom 1940 og 1959, og ble fulgt opp til og med 31.desember 2008. I 1990 ble utdannelsesnivået registrert, mens dødsårsakene ble registrert i perioden 1991 til 2008. Alderen til disse personene var altså mellom 32 og 68 år. Utdannelsesnivå ble delt inn etter antall år med fullført utdannelse. Inndelingen var slik:

- 7-9 år
- 10-11 år
- 12 år
- 12-16 år
- >16 år

Denne inndelingen er blitt brukt videre i oppgaven. Dødsårsakene ble delt inn i fem ulike grupper i tillegg til totaldødeligheten. Det er forskjellige koder for de fem ulike dødsårsakene. Kodesystemet som ble benyttet er *International Classification of Diseases*[1] (ICD-10), og er utviklet av Verdens Helseorganisasjon. Ved hjelp av dette systemet kan dødeligheten i forskjellige land sammenlignes, og en kan følge utviklingen av ulike dødsårsaker over tid.

[1] http://finnkode.kith.no/



Figur 2.1: *Illustrasjon av oppfølgingsperioden.*

Dødsårsakene og de tilhørende kodene er listet opp i Tabell 2.1. Videre i dette kapittelet er dødsårsakene beskrevet mer grundig.

Tabell 2.1: *Dødsårsaker og ICD-koder.*

Dødsårsak	**ICD-kode**
Hjerte-og karsykdommer (CVD, CHD)	I00-I99
Lungekreft	C32-C34
Alkoholrelaterte årsaker	F10, K70, K73-K74
Ulykker	V10-Y89

2.2 Variablene i datasettet

Som nevnt i innledningen, besto studien av 871 367 personer fordelt på 337 627 familier. Familier med kun ett barn ble utelatt fra analysene. Som nevnt, er datasettet som brukes i denne oppgaven er en annen versjon av det som ble brukt i studien og består av 505 043 familier. Dette svarer til totalt 934 548 norske barn som ble født 1940 og 1959, og av disse er 212 743 (22.8%) grupper på én person. Disse personene var ikke nødvendigvis enebarn. De kan ha hatt søsken som ble født før 1940 eller etter 1959 og som dermed ikke ble tatt med i studien. Vi kommer til å bruke betegnelsen ”enebarn” for disse personene gjennom hele oppgaven. I tillegg ble alle søskenflokker som besto av flere enn fire søsken fjernet fra datasettet, mens i artikkelen kunne søskenflokkene som bestå av flere enn fire søsken. Dette er gjort på grunn av personvernhensyn. Andelen kvinner og menn i dette datasettet er henholdsvis 48% (448 285) og 52% (486 263). I dette datasettet er det kun biologiske søsken, det vil si de hadde samme foreldre. Det er totalt 292 300 søskenflokker som består av to eller flere søsken. Andelen første- og andrefødte er omtrent likt, ca. 31%, og de utgjorde den største gruppen i datasettet. Dette kan vi se fra Tabell 2.2.

Tabell 2.2: *Antall "enebarn" og søsken.*

	Antall		Total
	Menn	**Kvinner**	**andel**
"Enebarn"	111 626	101 117	22.8%
Barn nr.1 i søskenflokken	154 191	136 977	31.1%
Barn nr.2 i søskenflokken	149 422	141 945	31.2%
Barn nr.3 i søskenflokken	56 272	53 913	11.8%
Barn nr.4 i søskenflokken	14 752	14 333	3.1%

For hver person er det fire forklaringsvariable. Disse variablene er fødselsår, kjønn, utdannelse og antall i husholdet. Tabellene 2.3, 2.4 og 2.5 viser henholdsvis oppsummering av variabelen fødselsår, hvordan de kategoriske variablene er kodet i datasettet og dødsårsaker. I tabellene A.1 og A.2 (Tillegg A) oppsummeres personene i datasettet delt inn etter fødselsår i tillegg til antall døde.

I tillegg til fødselsår i Tabell 2.3 ble også dødsår og dødsalder for personene registrert. Dødsårsakene ble registrert i tidsperioden 1991 til 2008. Det vil si at de personene som døde var mellom 32 og 68 år gamle.

Tabell 2.3: *Oppsummering av de kontinuerlige variablene.*

Variabelnavn	**Minimum**	**Median**	**Gjennomsnitt**	**Maksimum**
Fødselsår	1940	1947	1948	1959

Som nevnt i forrige seksjon, ble utdannelse delt inn i fem ulike grupper. Det første utdannelsesnivået er grunnskole. Grunnskole var sjuårig før 1969, men i 1969 ble den utvidet til niårig i Norge. Alle personene hadde dermed fullført minst sjuårig grunnskole. Det neste nivået, som er "10-11 år", kan defineres som yrkesskole. Som vi ser fra Tabell 2.4, gjaldt dette 30.6% av personene i dette datasettet. De utgjorde den største andelen. De som har gått 12 år på skolen har fullført gymnas/artium. Definisjonen "12-16 år" gjelder personene som har høyskoleutdannelse på lavere nivå, som for eksempel lærer- og sykepleierutdannelse, som er yrkesrettet utdannelse. Slike utdannelser har normal varighet på tre til fire år. De som har universitets- og høyskoleutdannelse på høyere nivå tilhører den siste kategorien, som varer i mist fire år.

Variabelen antall i husholdet angir antall personer i hver husholdning. Av personvernhensyn er denne variabelen trunkert ved fem. Som vi ser fra Tabell 2.4, er det 19% av husholdningene som består av fem personer eller flere.

Tabell 2.4: *Oppsummering av forklarings-variablene.*

Variabelnavn	Kode	Antall	Andel
Kjønn			
	Kvinne	448 285	47.9%
	Mann	486 263	52.1%
Utdannelse			
	2 = "7-9 år"	178 237	19.1%
	3 = "10-11 år"	286 433	30.6%
	4 = "12 år"	216 178	23.1%
	5 = "12-16 år"	114 324	12.3%
	6 = ">16 år"	124 624	13.3%
	99 = "Missing"	14 752	1.6%
Antall i husholdet			
	1	140 726	15.1%
	2	109 170	11.7%
	3	173 017	18.5%
	4	331 464	35.5%
	5 eller høyere	180 171	19.2%

I Tabell 2.5 er det en variabel for hver av de fem dødsårsakene. Variabelen har verdien 1 hvis en person døde av en av de nevnte dødsårsakene og 0 ellers. Her blir død av de andre dødsårsakene regnet som sensurering når vi ser på hver dødsårsak for seg. CVD (*Cardiovascular disease*) er hjerte- og karsykdommer, og er fellesbetegnelse for hjerteinfarkt, slag og andre hjertesykdommer. Dette er den vanligste dødsårsaken i Norge[2]. CHD (*Coronary heart disease*) er hjerteinfarkt og er inkludert i CVD. Forskning viser at tobakk- og sigarettrøyking er årsaken til de aller fleste tilfellene av lungekreft. Andre eksempler på årsaker til lungekreft er passiv røyking, luftforurensning og eksponering for radon og asbest[3]. Under dødsårsaken alkohol er blant annet leversykdommer og leversvikt. Eksempler på ulykker er transportulykker, ulykker med brann eller drukning.

[2]http://www.ssb.no/dodsarsak/
[3]https://kreftforeningen.no/om-kreft/kreftformer/lungekreft/

Tabell 2.5: *Oppsummering av respons-variablene.*

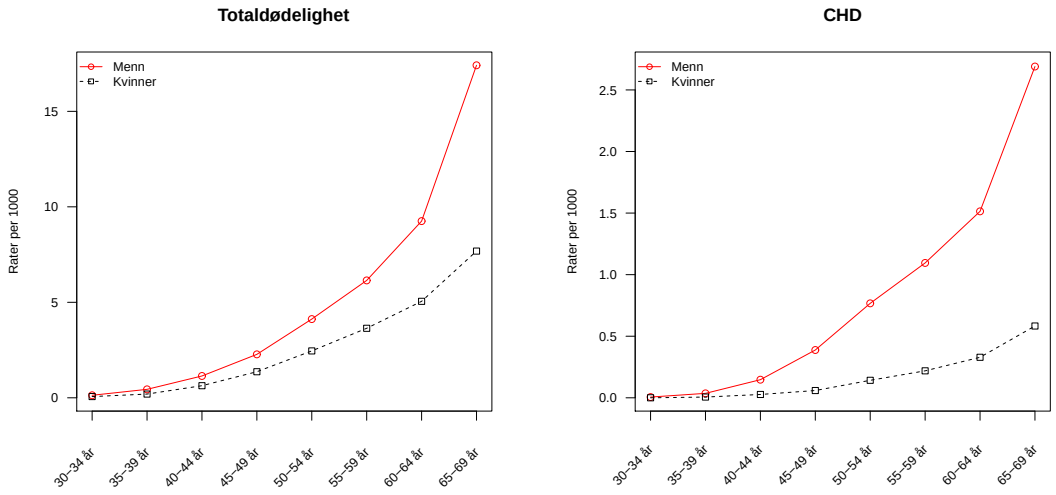
Variabelnavn	Kode	Antall	Andel
Lungekreft			
	1 = død pga lungekreft	4 515	0.5%
	0 = ikke død	930 033	99.5%
Alkohol			
	1 = død pga alkoholrelaterte årsaker	2 440	0.3%
	0 = ikke død	932 108	99.7%
CVD			
	1 = død pga CVD	9 579	1.0%
	0 = ikke død	924 969	99.0%
CHD			
	1 = død pga CHD	6 359	0.7%
	0 = ikke død	928 189	99.3%
Ulykker			
	1 = død pga ulykker	3 620	0.4%
	0 = ikke død	930 928	99.6%

2.3 Rater

Figur 2.2 viser de plottede ratene for totaldødelighet og CHD for ulike aldersgrupper. Tabell A.3 og Tabell A.4 i Tillegg A.2 viser de tilsvarende tallene. Vi ser at ratene for totaldødelighet er betydelig lavere for kvinner enn menn. Dette gjelder spesielt den siste aldergruppen; "65-69 år". Ratene for kvinner og menn i denne aldersgruppen ble henholdsvis 7.7 og 14.4. Fra Figur 2.3 ser vi antall døde for de tilsvarende gruppene. Tabell A.4 viser en oversikt over antall døde kvinner og menn og rater delt inn etter alder for dødsårsaken CHD. Det var totalt 6 359 personer som døde på grunn av CHD (hjerteinfarkt), og av disse var 84% menn. Høyt kolesterolnivå og røyking er de viktigste risikofaktorene for hjerteinfarkt[4]. I tillegg er kjønn, arv og alder også viktige risikofaktorer. Fra denne tabellen ser vi at det var betydelig flere menn enn kvinner som døde av hjerteinfarkt. Dette gjelder spesielt for menn i aldersgruppen "50-59 år". Det er stor forskjell mellom ratene for kvinner og menn. De tilsvarende tabellene for lungekreft og alkoholrelaterte årsaker er i Tillegg A.2. Lungekreft og alkoholrelaterte årsaker er hyppigere blant menn enn kvinner, men ikke i like stor grad som CHD. For begge disse dødsårsakene var det flest i aldersgruppen "50-59 år" som døde. Totalt var det henholdsvis 4 515 og 2 440 tilfeller av lungekreft og alkoholrelaterte årsaker. Som Tabell A.5 viser, øker ratene med alder, og forskjellen mellom ratene for kvinner og menn blir større. For lungekreft er det ikke så stor forskjell for de tre første aldersgruppene. For aldersgruppen ”45-49 år” er raten høyere for

[4] www.fhi.no/artikler/?id=70806

menn og dette gjelder også for de neste aldersgruppene. Også her er den største forskjellen for den siste gruppen; ”65-69 år”. For totaldødeligheten og de dødsårsakene som er nevnt over, er det flest dødsfall i aldersgruppen "50-59 år", mens ratene øker med alderen og er størst for den siste gruppen. Grunnen til det er at disse personene vil bidra med færre år sammenlignet med de i gruppen "50-59 år".



(a) (b)

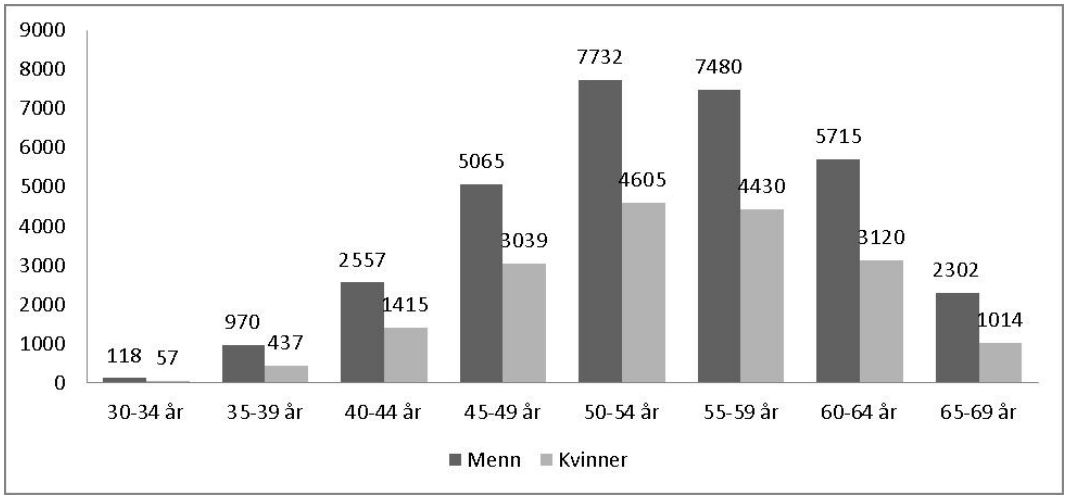
Figur 2.2: *Rater per 1000.*

2.4 Antall døde delt inn etter utdannelsesnivå

2.4.1 Hele datasettet

Tabell A.2 i Tillegg A angir antall døde kvinner og menn delt inn etter året de ble født. De flest søm døde var født mellom 1940 og 1949. Dette gjelder begge kjønn. Den totale andelen blant døde i dette tidsrommet var 64%. De fleste var mellom 50 og 59 år. Det var totalt 15 212 menn og 9 035 kvinner som døde i denne aldersgruppen. Figur 2.3 illustrerer dette.

Fra Tabell 2.6 ser vi også at antall døde minker når antall år med utdannelse øker. For menn gjelder dette for alle dødsårsakene. Det samme gjelder for kvinner, bortsett fra CVD og ulykker. For disse dødsårsakene var det flest som hadde fullført yrkesskole. Andelen dødsfall grunnet lungekreft, CVD, CHD, alkoholrelaterte årsaker og ulykker er henholdsvis 9.0%, 19.1%, 12.7%, 4.9% og 7.2%.



Figur 2.3: *Totalt antall døde.*

Tabell 2.6: *Antall døde med prosenter i parentes. Tallene i Tabell 2.4 er brukt her. Deler antall døde på antall i hver utdannelsesgruppe separat for kvinner og menn.*

Dødsårsak	7-9 år	10-11 år	12 år	12-16 år	>16 år	Missing	Totalt
Tot.død							
Menn	9 499 (10.5)	8 414 (7.5)	7 882 (5.4)	2 600 (4.5)	2 564 (3.5)	980 (12.9)	31 939
Kvinner	5 452 (6.2)	6 945 (4.0)	2 251 (3.2)	1 652 (2.9)	1 236 (2.2)	581 (8.1)	18 117
Lungekreft							
Menn	939 (1.0)	708 (0.6)	673 (0.5)	209 (0.4)	126 (0.2)	41 (0.5)	2 696
Kvinner	743 (0.8)	682 (0.4)	189 (0.7)	94 (0.2)	69 (0.2)	42 (0.7)	1 819
CVD							
Menn	2 404 (2.6)	1 983 (1.8)	1 822 (1.3)	558 (1.0)	521 (0.7)	199 (2.6)	7 487
Kvinner	779 (0.8)	816 (0.5)	235 (0.3)	125 (0.2)	72 (0.1)	65 (0.9)	2 092
CHD							
Menn	1 759 (1.9)	1 402 (1.3)	1 299 (0.9)	394 (0.7)	358 (0.5)	127 (1.7)	5 339
Kvinner	443 (0.5)	378 (0.2)	99 (1.4)	41 (0.1)	28 (0.1)	31 (0.4)	1 020
Alkohol							
Menn	694 (0.8)	545 (0.4)	417 (0.3)	113 (0.2)	93 (0.2)	60 (0.8)	1 922
Kvinner	214 (0.2)	195 (0.1)	51 (0.1)	24 (0.0)	16 (0.0)	18 (0.3)	518
Ulykker							
Menn	846 (0.9)	761 (0.7)	703 (0.5)	187 (0.3)	207 (0.3)	78 (1.0)	2 782
Kvinner	261 (0.3)	303 (0.2)	112 (0.2)	68 (0.1)	63 (0.1)	31 (0.4)	838

2.4.2 Tvillinger

Andelen flerfødsler i datasettet er 1.7%. Dette svarer til 15 796 personer. Av disse var andelen kvinner og menn henholdsvis 49.3% (7 781) og 50.7% (8 015). Her er det ikke skilt mellom eneggede og toeggede tvillinger.

Når det gjelder dødsårsakene, var det som for resten av datasettet CVD (hjerte-og kar-sykdommer) som forårsaket flest dødsfall for begge kjønn. Menn har, som for resten av datasettet, høyere dødelighet. Totalt var det 149 tvillinger som døde av CVD, og 75.8% av de var menn.

Tabell 2.7: *Antall døde tvillinger fordelt etter ant. menn (M) og kvinner (K).*

	Tot.død		**Lungekreft**		**CVD**		**CHD**		**Alkohol**		**Ulykker**	
Utdannelse	**M**	**K**	**M**	**K**	**M**	**K**	**M**	**K**	**M**	**K**	**M**	**K**
7-9 år	158	89	22	10	35	13	30	6	9	5	20	3
10-11 år	124	117	9	16	29	15	22	6	2	1	16	8
12 år	141	22	9	1	37	3	23	1	6	0	13	3
12-16 år	33	27	5	3	6	3	4	1	0	0	5	1
>16 år	30	23	2	1	4	0	3	0	3	0	1	3
Missing	15	7	0	0	2	2	1	0	2	0	2	0
Totalt	501	285	47	31	113	36	83	14	22	6	47	18

2.4.3 "Enebarn"

Som nevnt i innledningen, består 22.8% (212 743) av datasettet av "enebarn", altså grupper som besto av kun én person. Disse personene er ikke nødvendigvis enebarn, men kan ha søsken som er født før studien startet i 1940 eller etter 1959. Blant "enebarn" var det 12 942 (36.3% kvinner og 63.7% menn) som døde. Som Tabell 2.8 viser, er det igjen CVD og CHD som har forårsaket flest dødsfall. Til sammen er det 4 293 som har dødd på grunn av disse dødsårsakene, og 80.2% av de personene var menn.

Tabell 2.8: *Antall døde "enebarn". Andelen er antall døde delt på totalt antall døde.*

Dødsårsak	**Antall døde**	**Andel**
Totaldødelighet	12 942	
Lungekreft	1 150	8.9%
CVD	2 599	20.1%
CHD	1 694	13.1%
Alkohol	706	5.5%
Ulykker	850	6.6%

Kapittel 3

Teori og metoder

3.1 Levetidsanalyse

3.1.1 Sensurering

I levetidsanalyse er man interessert i tiden det tar til en hendelse inntreffer, for eksempel tid fra fødsel til død. Sensurering, det vil si ufullstendige observasjoner, er vanlig i levetidsanalyse. Det er flere årsaker til sensurering, for eksempel at studien avsluttes før hendelsen som er av interesse inntreffer for alle individene. Andre årsaker kan være at en person ikke lenger ønsker å delta i studien eller av ulike årsaker ikke kan følges opp lenger. Det finnes ulike typer sensurering; *høyre-, venstre- og intervallsensurering*. Den vanligste formen for sensurering er høyresensurering; når studien avsluttes før hendelsen har inntruffet. Siden vi bare kan observere levetidene inntil et tidspunkt, og mange av individene fortsatt er i live når oppfølgingstiden er over, vil levetidene bli høyresensurerte.

Vi lar T^* være tiden til hendelsen av interesse inntreffer. Dette er en tilfeldig variabel større enn 0. Observert levetid betegnes med $T = \min(T^*, C)$, der C er sensureringstiden. Vi lar δ være indikatorvariabelen for sensurering;

$$\delta = \begin{cases} 1 & \text{hvis } T^* \leq C, \\ 0 & \text{hvis } T^* > C. \end{cases}$$

Dette er en tilfeldig variabel som har verdi lik 1 hvis hendelsen inntreffer i løpet av studien, det vil si vi observerer T^*, og $\delta = 0$ hvis levetiden er sensurert. En viktig forutsetning er at sensurering er uavhengig av levetiden. Gitt kovariatene er T^* og C uavhengige, det vil si $T^*|\boldsymbol{x}$ er uavhengig av $C|\boldsymbol{x}$, der $\boldsymbol{x}$ er kovariater.

3.1.2 Venstre-trunkering

Når vi kun observerer de individene som har levetid større enn en nedre grense Y_L, det vil si $T^* > Y_L$, kalles det *venstre-trunkering*. Det er vanlig at dataene er høyresensurerte og

venstre-trunkerte. Siden individene kun observeres fra den tiden og fram til sensurering eller død, kalles denne formen for trunkering også *delayed entry*.

I datasettet som brukes i denne oppgaven, er levetidene høyresensurerte og venstre-trunkerte. Som beskrevet i kapitlene 1 og 2, er disse personene født mellom 1940 og 1959. Da studien ble avsluttet i 2008, levde 94.6%. Disse levetidene dermed høyresensurerte. I tillegg har vi venstre-trunkering; Y_L er alderen til personene i 1991 og varierer for alle individene. Laveste og høyeste trunkeringsalder vil være henholdsvis 32 og 51 år.

3.1.3 Hovedbegreper og resultater i levetidsanalyse med eksempler

Tettheten til T^* er

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* \leq t + \Delta t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}, \tag{3.1}$$

der $F(t) = P(T^* \leq t)$ er den *kumulative fordelingsfunksjonen. Overlevelsesfunksjonen* defineres som

$$S(t) = P(T^* > t) = 1 - F(t). \tag{3.2}$$

Dette er sannsynligheten for at levetiden blir minst t, og er den ubetingede sannsynligheten for at en hendelse ikke har inntruffet før denne tiden. Når T^* er en kontinuerlig tilfeldig variabel, er $S(t)$ en kontinuerlig, strengt minkende funksjon. Denne funksjonen starter i 1 og synker med tiden. Vi har også sammenhengen

$$f(t) = -S'(t). \tag{3.3}$$

Vi lar T_i^* være levetiden til individ i og observerer $T_i = \min(T_i^*, C_i)$. Vi antar at T_i^*-ene er uavhengig identisk fordelte med tetthet $f(t)$. $\delta_i = I(\{T_i^* \leq C_i\})$ er 0/1 indikatoren og har verdien 1 hvis T_i^* observeres og 0 hvis levetiden er sensurert. Det er da (T_i,δ_i) som observeres for hvert individ. Videre lar vi $t_{(i)}$ være i'te ordnede observerte hendelsestidspunkt, slik at $t_{(1)} < t_{(2)} < t_{(3)}, ...$

I tillegg har vi $Y_i(t) = I(\{\nu_i < t \leq T\})$, der ν_i er trunkeringstiden for individ i. Denne indikatoren er 1 hvis individet er under risiko, det vil si når individet har kommet under observasjon og ikke opplevd hendelsen innen tid t. Summen $Y(t) = \sum_{i=1}^{n} Y_i(t)$ er antall individer som er under risiko på tid t. Vi antar at to eller flere hendelser ikke kan skje på samme tid.

En vanlig estimator for overlevelsesfunksjonen er *Kaplan-Meier* estimatoren, og kan uttrykkes slik:

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{1}{Y(t_{(i)})}\right), \tag{3.4}$$

der $\hat{S}(t) = 1$ for $t < t_{(i)}$.

Hazardfunksjonen kan defineres slik:

$$h(t) = \frac{\lim\limits_{\Delta t \to 0} \dfrac{P(t \leq T^* \leq t + \Delta t)}{\Delta t}}{P(T^* > t)} = \lim_{\Delta t \to 0} \frac{P(T^* \leq t + \Delta t | T^* > t)}{\Delta t}. \tag{3.5}$$

Fra (3.5) ser vi at $h(t)\Delta t$ er den tilnærmede sannsynligheten for at hendelsen vil skje i intervallet $[t, t + \Delta t)$, gitt at hendelsen ikke har inntruffet innen tid t. I motsetning til $S(t)$, som alltid starter i 1, kan hazardfunksjonen være en vilkårlig ikke-negativ funksjon. Hazardraten kan også skrives som

$$h(t) = \frac{f(t)}{S(t)}. \tag{3.6}$$

Integralet av (3.5) mellom 0 og t gir den *kumulative hazardfunksjonen* på tid t;

$$H(t) = \int_0^t h(u)du. \tag{3.7}$$

Sammenhengen mellom overlevelsesfunksjonen og hazardfunksjonen når levetidene er kontinuerlige er

$$S(t) = \mathrm{e}^{-H(t)}. \tag{3.8}$$

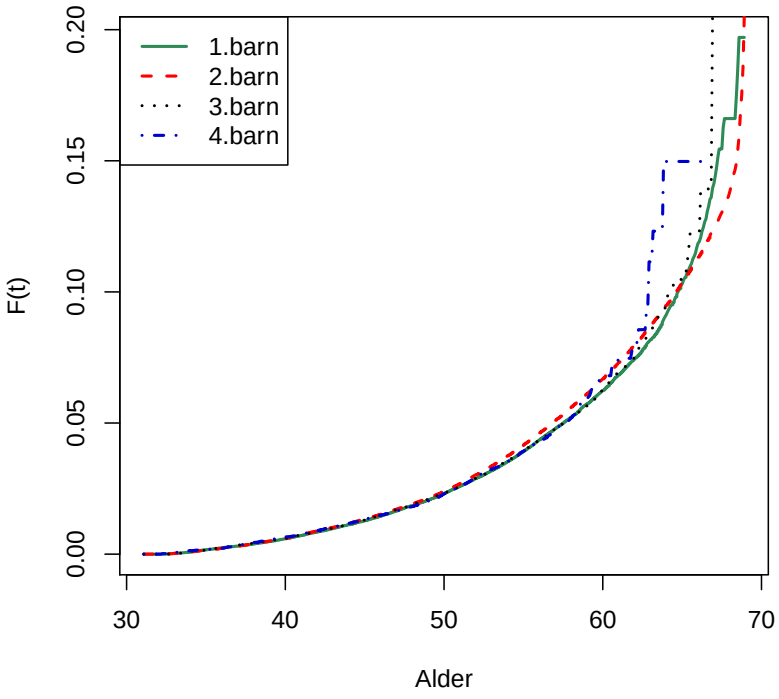
Denne sammenhengen viser at for å estimere $S(t)$, kan en estimator for $H(t)$ brukes, og motsatt. *Nelson-Aalen*-estimatoren

$$\hat{H}(t) = \sum_{t_{(i)} \leq t} \frac{1}{Y(t_{(i)})} \tag{3.9}$$

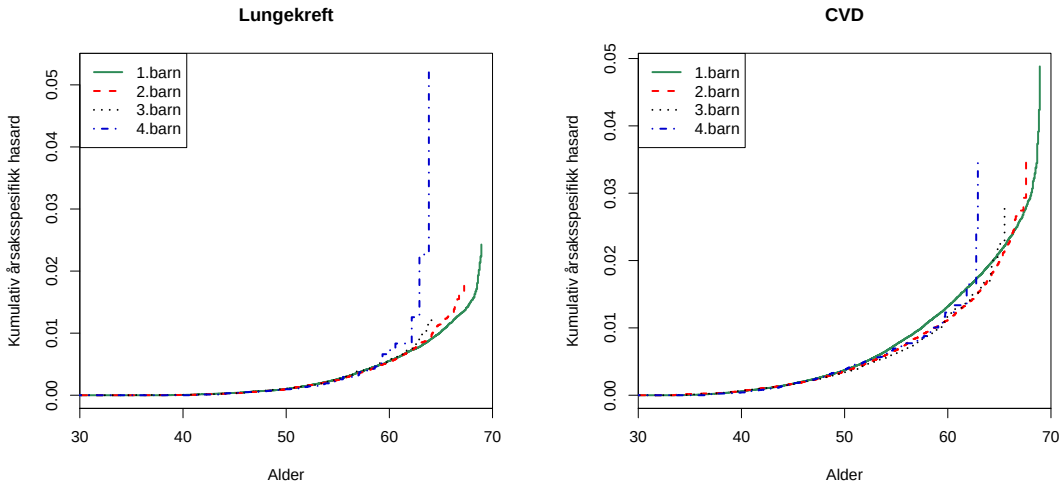
brukes til å estimere den kumulative hazardfunksjonen. Her er $Y(t)$ som nevnt over antall individer som er under risiko på tid t. Det er enklere å estimere den kumulative hazardfunksjonen enn $h(t)$ direkte.

Eksempel 3.1. *Kaplan-Meier og Nelson-Aalen.* Figur 3.1 viser 1 - $\hat{S}(t)$, som er den estimerte sannsynligheten for at et individ dør ved alder t. I denne figuren er det skilt mellom første-, andre, tredje og fjerdefødte i en søskenflokk, og her ser vi på totaldødeligheten. Overlevelsessannsynligheten er høy, og det er spiller ikke så stor rolle hvilket barn man er i rekken. Som figuren viser, er sannsynligheten for at første barn overlever til 68 år estimert til 80%.

Figur 3.2 viser plott av Nelson-Aalen estimater for dødsårsakene lungekreft og CVD for førstefødte, andrefødte osv. Fra denne figuren ser vi at det ikke er betydelig forskjell mellom Nelson-Aalen plottene for et barn i søskenflokken. Kurvene er voksende, og dermed øker hazard-raten med alder. I Tillegg B er de tilsvarende plottene for CHD, alkoholrelaterte årsaker og ulykker vist. For de andre dødsårsakene observerer vi det samme som for lungekreft og CVD, men for ulykker (Figur B.2) ser kurvene ut til å være mer lineære. Dette indikerer en tilnærmet konstant hazard.



Figur 3.1: *Totaldødelighet:* $\hat{F}(t) = 1 - \hat{S}(t)$



(a) (b)

Figur 3.2: *Nelson-Aalen*

3.2 Cox-modell

Forklaringsvariable som alder, kjønn, utdannelse og økonomi blir brukt til å forklare responsvariablene. Cox proporsjonale hazard modell for individ i er definert som produktet av to funksjoner;

$$h(t, \mathbf{X}_i, \boldsymbol{\beta}) = h_0(t)\exp\{\boldsymbol{\beta}'\mathbf{X}_i\}, \tag{3.10}$$

der $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ er en vektor som består av p regresjonskoeffisienter, mens $\mathbf{X}_i = (X_{i1}, ..., X_{ip})'$ er en vektor med p kovariater for individ i. Kovariatene kan være tidsavhengige, som for eksempel blodtrykk. Eller de kan også være konstante og kjent ved tid 0, som for eksempel kjønn. En av fordelene med denne modellen er at den tillater flere variable. Funksjonen $h_0(t)$ kalles *baseline* og er ikke-parametrisk i denne modellen. Her er det ikke nødvendig å spesifisere den. I denne modellen er hazardfunksjonen avhengig av kovariatene og $\exp\{\boldsymbol{\beta}'\mathbf{X}\}$, som er parametrisk, viser hvordan hazardfunksjonen forandres som en funksjon av kovariater. Når alle kovariatene er lik 0, blir $h(t, 0, \boldsymbol{\beta})$ baseline hazard.

3.2.1 Relativ risiko

For Cox-modeller brukes *hazard ratio* (HR), og er uttrykt ved eksponential opphøyd i en eller flere regresjonskoeffisienter, og avhenger ikke av tid. For tidsuavhengige kovariater

blir HR en konstant. Hazard ratioen for to individer med kovariater $\mathbf{X}_1$ og $\mathbf{X}_2$ er gitt ved

$$\frac{h(t,\mathbf{X}_2,\boldsymbol{\beta})}{h(t,\mathbf{X}_1,\boldsymbol{\beta})} = \frac{h_0(t)\exp\{\sum_{k=1}^{p}\beta_k X_{2k}\}}{h_0(t)\exp\{\sum_{k=1}^{p}\beta_k X_{1k}\}} \tag{3.11}$$

$$= \exp\left\{\sum_{k=1}^{p}\beta_k(X_{2k}-X_{1k})\right\}. \tag{3.12}$$

Spesielt blir (3.11) lik $\exp\{\beta_j\}$ hvis $X_{2j}=1$ og $X_{1j}=0$ og alle de andre kovariatene er like. $\log(\text{HR}) = \sum_{k=1}^{p}\beta_k(X_{2k}-X_{1k})$ er da forskjellen i *log-hazardfunksjonen*. Forskjellen i log-hazard avhenger heller ikke av tid under Modell (3.10).

Eksempel 3.2. I denne oppgaven er $i = 1, 2, ..., 934\ 548$, som er antall personer. Et eksempel på resultater fra en Cox regresjon i `R` er vist i Tabell 3.1. Her tilpasses en Cox-modell kun for totaldødelighet.

Tabell 3.1: *Antall i husholdet.*

Antall	$\mathbf{exp}(\hat{\boldsymbol{\beta}})$
2	0.65
3	0.51
4	0.37
5 eller høyere	0.36

Det er ikke justert for andre kovariater. Alle p-verdiene for dette eksempelet er tilnærmet 0, det vil si alle er signifikante. For denne variabelen ser vi at risikoen minker når antall personer i husholdet øker når 1 brukes som referanse. For eksempel er den estimerte risikoen for død ca. to ganger høyere for en person der husholdet består av én person sammenlignet med en person der det er tre i husholdet. Her er det gruppen med fem eller flere personer i husholdet som har minst risiko for død. Risikoen er omtrent 60% lavere.

3.2.2 Partiell likelihood

For å finne estimatene for parameterne bruker vi *partiell likelihood*. Vi lar $t_1 < t_2 < ... < t_D$ være de ordnede tidspunktene individene dør på, det vil si D er antall døde og $\Re(t)$ er antall individer under risiko rett før tidspunkt t. Det er de individene som ikke er sensurerte eller har opplevd hendelsen som er av interesse. Vi antar at to individer ikke kan dø på samme tidspunkt. De dataene som er venstre-trunkerte kommer inn under risiko når levetiden er større enn en øvre grense, og individene er under risiko fram til sensurering eller død. Cox-modell tar automatisk hensyn til venstre-trunkering.

Vi lar $X_{(i)k}$ være kovariat k for individ i med levetid t_i. Den betingede sannsynligheten for at individ i dør på tid t gitt at en hendelse har inntruffet for et individ i risksettet

$\Re(t)$ er

$$L_i(\boldsymbol{\beta}) = \frac{h(t, \boldsymbol{X}_i, \boldsymbol{\beta})}{\sum_{j \in \Re(t_i)} h(t, \boldsymbol{X}_j, \boldsymbol{\beta})} = \frac{\exp\{\sum_{k=1}^{p} \beta_k X_{(i)k}\}}{\sum_{j \in \Re(t_i)} \exp\{\sum_{k=1}^{p} \beta_k X_{jk}\}}.$$

Dette er likelihood-bidraget fra individ i. Partiell likelihood over antall døde blir

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\{\sum_{k=1}^{p} \beta_k X_{(i)k}\}}{\sum_{j \in \Re(t_i)} \exp\{\sum_{k=1}^{p} \beta_k X_{jk}\}}. \tag{3.13}$$

I Formel (3.13) forkortes $h_0(t)$ bort, og $L_i(\boldsymbol{\beta})$ avhenger bare av regresjonsparameteren $\boldsymbol{\beta}$. En estimator for $\boldsymbol{\beta}$ finnes ved å maksimere (3.13) som vanlig likelihood. Man kan altså estimere $\boldsymbol{\beta}$ uten å spesifisere baseline hazard.

Log-likelihood blir

$$\begin{aligned} \log[L(\boldsymbol{\beta})] &= \sum_{i=1}^{D} \log\left[\frac{\exp\{\sum_{k=1}^{p} \beta_k X_{(i)k}\}}{\sum_{j \in \Re(t_i)} \exp\{\sum_{k=1}^{p} \beta_k X_{jk}\}}\right] \\ &= \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k X_{(i)k} - \sum_{k=1}^{D} \log\left[\sum_{j \in \Re(t_i)} \exp\{\sum_{k=1}^{p} \beta_k X_{jk}\}\right]. \end{aligned} \tag{3.14}$$

Man kan også finne likelihood estimatene ved å maksimere (3.14). Ved å derivere (3.14) med hensyn på $\boldsymbol{\beta}$, får man vektoren med *scorefunksjoner*

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log[L(\boldsymbol{\beta})].$$

Videre har vi *observert informasjon*

$$I(\boldsymbol{\beta}) = -U'(\boldsymbol{\beta}) = [I_{gh}(\boldsymbol{\beta})]_{p \times p}, \tag{3.15}$$

som er en $p \times p$-matrise. Når individene er uavhengige har vi denne sammenhengen:

$$\mathrm{Var}[U(\boldsymbol{\beta})] = \mathrm{E}[I(\boldsymbol{\beta})]. \tag{3.16}$$

Man kan vise at $L(\boldsymbol{\beta})$ kan benyttes som en regulær likelihood, og at $\hat{\boldsymbol{\beta}}$ er tilnærmet normalfordelt med disse egenskapene :

- $\mathrm{E}[\hat{\boldsymbol{\beta}}] \approx \boldsymbol{\beta}$
- Estimert kovariansmatrise $[I(\hat{\boldsymbol{\beta}})]^{-1}$.

Hittil har vi antatt at alle individene er uavhengige av hverandre. For datasettet i denne oppgaven vil det si at det ikke er noe avhengighet mellom personene. Det er ikke realistisk, fordi mange av de er søsken. Videre skal vi bruke modeller som tar hensyn til familieavhengigheten. Dette gjøres i Seksjon 3.2.4 og Seksjon 3.4.

3.2.3 Konfidensintervaller for HR

Fra generell teori for konfidensintervaller har vi at et tilnærmet 100(1-α)% konfidensintervall for HR_j er

$$\widehat{HR}_j \exp\{\pm z_{1-\alpha/2}\text{se}(\hat{\beta}_j)\},$$

der $\text{se}(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)}$ er standardfeil for $\hat{\beta}_j$ og er element j på diagonalen i $I(\hat{\boldsymbol{\beta}})^{-1}$. Utledningen kan gjøres ved å starte med et $100(1-\alpha)\%$ konfidensintervall for β_j. Dersom intervallet dekker 1, indikerer det at hazardratene for to grupper er like.

Tabell 3.2: *Konfidensintervaller fo HR.*

Antall	**95% konf.int**
2	(0.63 , 0.66)
3	(0.50 , 0.52)
4	(0.37 , 0.39)
5 eller høyere	(0.34 , 0.37)

Eksempelet i Tabell 3.2 er en fortsettelse av Eksempel 3.2. Denne tabellen viser 95% konfidensintervaller for hazard ratioen. I dette tilfellet er det ingen av konfidensintervallene som inneholder verdien 1.

3.2.4 Marginal modell for multivariate data

En alternativ modell for multivariate data er marginal proporsjonal hazard modell

$$h_{ij}(t, \boldsymbol{X}_{ij}, \boldsymbol{\beta}) = h_0(t)\exp\{\boldsymbol{\beta}' \boldsymbol{X}_{ij}\} \quad \text{for} \quad i = 1, 2, ..., n_j \quad \text{og} \quad j = 1, 2, ..., G. \tag{3.17}$$

I Modell (3.17) er $h_{ij}(t, \boldsymbol{X}_{ij}, \boldsymbol{\beta})$ hazardraten for individ i i gruppe j. Vi antar at det er avhengighet mellom individene i hver av de G gruppene.

For å estimere $\boldsymbol{\beta}$, bruker man samme fremgangsmåte som for Cox-modell der vi maksimerer partiell likelihood (3.13). Vi antar uavhengighet mellom dataene og har $\sum_{j=1}^{G} n_j$ observasjoner, men varians-estimatoren blir ikke den samme. For å estimere variansen, brukes en *sandwich-estimator*. Denne estimatoren tar hensyn til mulig avhengighet mellom gruppene, og kan skrives som:

$$\widehat{\text{Var}}(\boldsymbol{\beta}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \tag{3.18}$$

der $\mathbf{B} = \sum_i \mathbf{U}_i\mathbf{U}_i^T$ er en sum over alle individene, mens $\mathbf{A}^{-1}$ er informasjonsmatrisen til $\hat{\boldsymbol{\beta}}$. Her er $\mathbf{U}_i$ individuelle score-residualer[1].

[1] Modeling Survival Data: Extending the Cox Model, s.173

3.3 Stratifisert Cox-modell

I Cox-modell (3.10) antar vi at baseline hazard er lik for alle individene. I den stratifiserte modellen (3.19) har hver gruppe en baseline hazard. Baseline hazard kan utvikle seg uavhengig over tid for hver gruppe. Hvis individ i tilhører gruppe j og S er antall strata, blir hazardraten

$$h_j(t, \boldsymbol{X}, \boldsymbol{\beta}) = h_{0j}(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}) \quad \text{for} \quad j = 1, 2, ..., S. \tag{3.19}$$

Tolkningen av hazardraten er som før. For å estimere $\boldsymbol{\beta}$, bruker man den partielle likelihooden fra stratum s:

$$L_{(s)}(\boldsymbol{\beta}) = \prod_{i=1}^{D_s} \left[\frac{\exp\{\sum_{k=1}^{p} \beta_{ks} X_{(i)ks}\}}{\sum_{j \in \Re(t_{si})} \exp\{\sum_{k=1}^{p} \beta_{ks} X_{jks}\}} \right]^{\delta_{is}}. \tag{3.20}$$

Her er t_{si} er tiden for individ i og *stratum* s og $\Re(t_{si})$ antall som er under risiko på dette tidspunktet. D_s er antall døde i stratum s. δ_{il} er en indikator som er lik 1 hvis individ i tilhører stratum s og 0 ellers. $\boldsymbol{\beta}$ er som før en vektor som består av p regresjonskoeffisienter, mens X_{si} er en vektor med p kovariater.

Produktet av partielle likelihood over hvert stratum blir

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{S} L_{(s)}(\boldsymbol{\beta}), \tag{3.21}$$

der $\hat{\boldsymbol{\beta}}$ har de samme egenskapene som for modellen uten stratifisering. I denne oppgaven svarer j til antall familier. Som nevnt i Kapittel 2, er det 292 300 søskenflokker i datasettet som brukes her. Dette svarer til 721 805 personer.

3.4 Univariat frailty

En utvidelse av Cox-modell (3.10) er

$$h_f(t, \boldsymbol{X}, \boldsymbol{\beta}) = Zh(t, \boldsymbol{X}, \boldsymbol{\beta}). \tag{3.22}$$

Dette er hazardraten multiplisert med Z. Her er $h_f(t, \boldsymbol{X}, \boldsymbol{\beta})$ hazardfunksjonen for et individ på tid t gitt en *frailty*[2]-variabel Z. Dette er en tilfeldig variabel som ikke kan observeres. Jo høyere verdi av Z, desto høyere er risikoen. De individene som er mest "skrøpelige" vil ha større sannsynlighet for å dø tidlig enn de andre. Dette vil føre til at de som er igjen under observasjon vil ha lavere gjennomsnittlig frailty. Hvis Z er mindre enn 1, er individet mindre "skrøpelige" enn gjennomsnittet, og motsatt hvis denne verdien er større enn 1. Det finnes flere alternativer for fordelingen til Z, for eksempel *log-normal frailty-modell* eller *gamma frailty-modell*. Det er ofte slik at hazardfunksjonen øker i starten og når maksimum før den avtar.

[2]På norsk oversettes dette ordet med "skrøpelig".

For en gitt Z er den betingede overlevelsessannsynligheten på tid t gitt som

$$S(t|Z) = P(t > T|Z) = \mathrm{e}^{-ZH(t)}, \tag{3.23}$$

der $H(t)$ er den kumulative hazardfunksjonen. Dette er på individnivå og kan ikke observeres. I tillegg kan hazardraten for populasjonen ha ulik form enn hazardraten på individnivå.

Ved å bruke regelen for dobbel forventning finner vi at overlevelsesfunksjonen på populasjonsnivå er gitt ved

$$S(t) = \mathrm{E}[S(t|Z)]. \tag{3.24}$$

Dette er overlevelsessannsynligheten for en tilfeldig valgt person.

En viktig sammenheng som brukes er *Laplace*-transformasjonen av Z:

$$\mathcal{L}(c) = \mathrm{E}(\mathrm{e}^{-cZ}). \tag{3.25}$$

Ved å bruke Formel (3.25), kan vi skrive $S(t)$ som

$$S(t) = \mathrm{E}(\mathrm{e}^{-ZH(t)}) = \mathcal{L}(H(t)). \tag{3.26}$$

Dermed blir hazard raten for populasjonen lik

$$\mu(t) = h(t)\frac{-\mathcal{L}'(H(t))}{\mathcal{L}(H(t))}. \tag{3.27}$$

Når vi ikke har frailty-effekt, blir hazard-funksjonen for populasjonen lik hazarden på individnivå.

3.4.1 Gamma frailty

Det er vanlig å anta at frailty-variabelen er gammafordelt. Tettheten er

$$f_Z(z) = \frac{z^{\gamma-1}\exp(-z/\theta)}{\theta^{\gamma}\Gamma(\gamma)} \quad \text{for } Z, \gamma, \theta > 0, \tag{3.28}$$

der γ og θ er to positive parametre. Vi setter $\theta = 1/\gamma$ slik at forventningen blir 1, og dermed er $Z \sim \Gamma(\theta, \theta)$. Ved å bruke denne parametriseringen blir tettheten

$$f_Z(z) = \frac{z^{1/\theta-1}\exp(-z/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)}. \tag{3.29}$$

Vi får da at $\mathrm{E}(Z) = 1$ og $\mathrm{Var}(Z) = \sigma^2 = \theta = \frac{1}{\gamma}$.

Variansen til Z, θ, sier noe om variasjonen i populasjonen. Jo større θ er, jo høyere er variasjonen i populasjonen. Det betyr større avhengighet innad i gruppene. Laplace-transformasjonen er

$$\mathcal{L}(c) = \left(1 + \frac{1}{\gamma}c\right)^{-1/\gamma}. \tag{3.30}$$

Formel (3.30) gir at overlevelsesfunksjonen og populasjons-hazarden blir henholdsvis

$$S(t) = [1 + \gamma H(t)]^{-1/\gamma} \tag{3.31}$$

og

$$\mu(t) = \frac{h(t)}{1 + \gamma H(t)}. \tag{3.32}$$

3.4.2 Log-normal frailty

Det finnes flere fordelinger for Z. En av dem er log-normal. Tettheten til log-normalfordelingen er

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma z}\exp\left(-\frac{(\log z - \mu)^2}{2\sigma^2}\right) \quad \text{for} \quad z > 0, \tag{3.33}$$

det vil si $\log(Z) \sim N(\mu, \sigma^2)$, og μ og σ^2 er parametre. Forventning og varians er henholdsvis lik

$$\mathrm{E}(Z) = \mathrm{e}^{\mu + \frac{1}{2}\sigma^2} \quad \text{og} \quad \mathrm{Var}(Z) = (\mathrm{E}(Z))^2\left(\mathrm{e}^{\sigma^2} - 1\right). \tag{3.34}$$

Ved å normere $\mathrm{E}(Z) = 1$, blir variansen til frailty-variabelen $\mathrm{e}^{\sigma^2} - 1$. I dette tilfellet kan ikke uttrykket for Laplace-transformasjonen i Formel (3.35) forenkles, og er gitt ved

$$\mathcal{L}(c) = \int_0^\infty \mathrm{e}^{-cZ} f_Z(z)dz. \tag{3.35}$$

3.5 Multivariat analyse

3.5.1 Multivariat frailty

I flere av analysene antar vi at levetidene til individene er uavhengige. Dette gjelder for eksempel ikke for søsken. For å kunne analysere slike data, kan man bruke frailty modeller. Frailty gjelder for individer, men også grupper. Det er en viss avhengighet innad i gruppene, og man antar at alle i gruppen har samme frailty. Frailty er en uobservert effekt som er lik for alle individene i en subgruppe. Den mest vanlige modellen for frailty er *shared frailty*. Modellen kalles shared frailty-modell fordi alle individene i samme gruppe deler samme frailty.

Shared frailty-modell

Shared frailty-modell er gitt ved

$$h_{ij}(t, \boldsymbol{X}_{ij}, \boldsymbol{\beta}) = Z_j h_0(t)\exp(\boldsymbol{\beta}'\mathbf{X}_{ij}) \quad \text{for} \quad i = 1, 2, ..., n_j \quad \text{og} \quad j = 1, 2, ..., G, \tag{3.36}$$

der hazardfunksjonen gitt frailty-variabelen Z_j. Dette er for gruppe j og individ i. $\boldsymbol{X}$ er som før en vektor med p kovaritater, og $\boldsymbol{\beta}$ er vektoren med regresjonskoeffisienter. Hver

gruppe har en Z_j, mens denne verdien er konstant over tid og lik for alle individene i samme gruppe. Det er avhengighet innad i gruppene, mens hver gruppe er uavhengig av hverandre. Z_j-ene er uavhengige og identisk fordelte.

Den marginale overlevelsessannsynligheten for gruppe j er

$$\begin{aligned} S_j(t_1, t_2, ..., t_{n_j}) &= P(T_1 > t_1, T_2 > t_2, ..., T_n > t_{n_j}) \\ &= \mathcal{L}(H_{j1}(t_{j1}) + H_{j2}(t_{j2}) + ... + H_{jn_j}(t_{jn_j})) \quad \text{for} \quad j = 1, 2, ..., G. \end{aligned} \tag{3.37}$$

3.6 Bootstrapping

I bootstapping ønsker man å estimere varians og skjevhet for en estimator $\hat{\theta}$ for en ukjent parameter θ. For å finne $\hat{\theta}$ bruker man observerte data $x_1, x_2, ..., x_n \sim F$. Sannsynlighetsfordelingen F er ukjent og estimeres ut fra observasjonene. I hovedsak skilles det mellom *parametrisk-* og *ikke-parametrisk bootstrapping.* I førstnevnte simulerer man fra en empirisk fordeling, som er beskrevet ved en eller flere parametre. Her vil vi kun fokusere på *ikke-parametrisk bootstrapping.*

3.6.1 Ikke-parametrisk bootstrapping

Siden fordelingen er til F er ukjent, defineres den som en diskret fordeling med mulige verdier gitt ved de opprinnelige observasjonene. Ideen er å trekke med tilbakelegging fra de observerte dataene til å generere nye datasett, $x_1^*, x_2^*, ..., x_n^*$, med sannsynlighet $1/n$ for hver observasjon. Denne prosedyren gjentas B ganger.

I denne oppgaven ønsker vi å estimere regresjonskoeffisienter, så $\theta = \beta$. En starter med å trekke tilfeldig 503 275 familier fra datasettet med tilbakelegging og deretter tilpasse en regresjonsmodell som gir $\hat{\beta}$. Deretter gjentas dette B ganger og får da et bootstrap-estimatet $\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b^*$. Ved å beregne dette gjennomsnittet, kan vi sammenligne det med $\hat{\beta}$ fra Cox regresjon. Hvis avviket er stort, er estimatoren forventningsskjev.

3.6.2 Konfidensintervaller

Når vi antar at $\hat{\theta}$ er tilnærmet normalfordelt, blir et tilnærmet 95% konfidensintervall for θ lik

$$\hat{\theta} \pm 1.96 * s_{\hat{\theta}}. \tag{3.38}$$

I Formel (3.38) er $s_{\hat{\theta}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\theta_b^* - \bar{\theta}^*)^2}$, som er standardfeilen til bootstrap-estimatene.

En annen måte å beregne konfidensintervaller på er å bruke persentilmetoden. I dette tilfellet definerer vi en generell θ. Vi har en vektorer med lengde B, som sorteres i stigende rekkefølge. Den nedre grensen i konfidensintervallet blir element nummer B*2.5%, mens den øvre grensen blir element nummer B*97.5%

3.6.3 Bootstrapping for testing av resultater

Vi kan bruke persentilmetoden til å teste om det er systematiske forskjeller mellom tilpasset Cox regresjonsmodell med og uten stratifisering på søskenflokk. I dette tilfellet definerer vi θ fra forrige seksjon til å være $\hat{\boldsymbol{\beta}}_{Full}$-$\hat{\boldsymbol{\beta}}_{Stratifisert}$ og finner et konfidensintervall for denne forskjellen. Vektorene $\hat{\boldsymbol{\beta}}_{Full}$ og $\hat{\boldsymbol{\beta}}_{Stratifisert}$ er hver av lengde B. Etter å ha funnet differansen mellom de to vektorene, sorteres elementene i denne vektoren i stigende rekkefølge. Den nedre grensen i konfidensintervallet blir element nummer B*2.5%, mens den øvre grensen blir element nummer B*97.5%. I dette tilfellet har vi avhengige data. Dersom konfidensintervallet inneholder 0, vil det si at det ikke er en signifikant forskjell på nivå 5%. Hvis det ikke er tilfellet, tyder det på at søskenavhengigheten er viktig her.

Kapittel 4

Resultater for hele datasettet

I dette kapittelet skal vi undersøke om søskenavhengigheten har betydning for dødsrisiko. For å gjøre det skal vi sammenligne flere metoder som tar hensyn til avhengighet mellom søsken på ulike måter. Vi skal først tilpasse en Cox regresjonsmodell uten stratifisering der alle individene i datasettet regnes som uavhengige i tillegg til marginal Cox-modell i Kapittel 3.2.4 som tar hensyn til familieavhengighet. Deretter tilpasses en Cox regresjonsmodell med stratifisering på søskenflokk. Videre er det resultater fra en frailty-analyse. I tillegg sammenlignes de ulike metodene, der vi konsentrerer oss om noen av dødsårsakene. Dette er gjort i Kapittel 4.4, Kapittel 4.5, Tillegg C.4 og Tillegg C.5. Grunnen til at vi har valgt å fokusere på disse dødsårsakene er at de hadde de høyeste hazard ratioene i Tabell 4.1. I neste kapittel skal vi bruke bootstrapping for å teste resultatene vi har fått. Vi skal sammenligne gjennomsnittene fra bootstraping med resultatene i dette kapittelet. Her har vi har stort sett brukt hele datasettet, det vil si både "enebarn" og søsken med unntak av Kapittel 4.6 der vi sammenligner resultatene for "enebarn" og søsken.

4.1 Resultater fra tilpasset Cox regresjonsmodell uten stratifisering

Tabell 4.1 viser estimert hazard ratio og tilsvarende 95% konfidensintervall for hazard ratio for alle dødsårsakene i tillegg til totaldødeligheten. Det er gjort separate analyser for kvinner og menn. Metoden som er brukt er Cox regresjonsmodell uten stratifisering. I det første tilfellet antar vi at alle de 934 548 personene i datasettet er uavhengige av hverandre. Dette er i utgangspunktet ikke realistisk, men gjøres for å se hvor viktig søskenavhengigheten er. I det andre tilfellet (*cluster*) brukes marginal Cox-modell som tar hensyn til familieavhengighet mellom søsken. Begge metodene gir de samme estimatene for hazard ratioen, men ulike konfidensintervaller på grunn av ulike varians-estimater. Her blir utdannelse på over 16 år brukt som referansegruppe. Totalt var det 124 624 (13.3%) personer i denne gruppen.

Tabell 4.1: *Cox regresjon uten stratifisering.*

		Kvinner			Menn		
			95% k.i			95% k.i	
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**Cox**	**Cluster**	$\widehat{\mathbf{HR}}$	**Cox**	**Cluster**
Total-dødelighet							
	7-9 år	2.19	2.06-2.33	2.06-2.33	2.87	2.75-3.00	2.75-3.00
	10-11 år	1.51	1.42-1.60	1.42-1.60	2.23	2.13-2.33	2.13-2.33
	12 år	1.37	1.28-1.47	1.28-1.47	1.66	1.59-1.74	1.59-1.74
	12-16 år	1.14	1.06-1.23	1.06-1.23	1.32	1.25-1.40	1.25-1.40
	>16 år	Ref.			Ref.		
Lungekreft							
	7-9 år	5.01	3.92-6.42	3.92-6.42	5.53	4.59-6.66	4.58-6.67
	10-11 år	2.55	1.99-3.27	1.99-3.27	3.85	3.18-4.65	3.18-4.66
	12 år	2.11	1.60-2.78	1.60-2.78	2.97	2.46-3.59	2.45-3.60
	12-16 år	1.15	0.84-1.56	0.84-1.56	2.18	1.75-2.72	1.74-2.72
	>16 år	Ref.			Ref.		
CVD							
	7-9 år	5.26	4.13-6.70	4.13-6.71	3.54	3.22-3.90	3.22-3.90
	10-11 år	3.00	2.36-3.82	2.36-3.82	2.59	2.35-2.85	2.35-2.85
	12 år	2.48	1.90-3.23	1.90-3.22	1.91	1.73-2.10	1.73-2.10
	12-16 år	1.48	1.11-1.97	1.11-1.97	1.40	1.24-1.58	1.24-1.58
	>16 år	Ref.			Ref.		
CHD							
	7-9 år	7.60	5.19-11.14	5.19- 11.15	3.79	3.39-4.25	3.39-4.25
	10-11 år	3.55	2.42-5.21	2.42-5.22	2.66	2.37-2.99	2.37-2.99
	12 år	2.70	1.77-4.10	1.77-4.10	1.98	1.76-2.22	1.76-2.22
	12-16 år	1.24	0.77-2.01	0.77- 2.01	1.44	1.25-1.66	1.25-1.66
	>16 år	Ref.			Ref.		
Alkoholrelaterte årsaker							
	7-9 år	6.91	4.19-11.50	4.15-11.51	5.95	4.79-7.39	4.79-7.39
	10-11 år	3.32	2.00-5.53	1.99-5.53	3.96	3.18-4.93	3.18-4.93
	12 år	2.37	1.35-4.16	1.35-4.16	2.40	1.92-3.00	1.92-3.00
	12-16 år	1.29	0.68-2.43	0.68-2.43	1.58	1.20-2.08	1.20-2.08
	>16 år	Ref.			Ref.		
Ulykker							
	7-9 år	2.27	1.72-2.99	1.72- 3.00	3.36	2.89-3.91	2.89-3.91
	10-11 år	1.35	1.03-1.78	1.03-1.78	2.46	2.11-2.87	2.11-2.87
	12 år	1.30	0.96-1.77	0.96-1.77	1.76	1.51-2.06	1.51-2.06
	12-16 år	0.95	0.67-1.33	0.67-1.33	1.17	0.96-1.42	0.96-1.42
	>16 år	Ref.			Ref.		

Ved å sammenligne 95% konfidensintervaller for Cox-modell og marginal Cox-modell i Tabell 4.1, ser vi ser at det er svært liten forskjell mellom konfidensintervallene for disse metodene. For den sistnevnte metoden bruker vi *sandwich-estimator*. For de dødsårsakene det er noe forskjell, som for eksempel CVD for kvinner og lungekreft for menn, er konfidensintervallene noe bredere for marginal Cox-modell, men disse forskjellene er ubetydelig små. Det er altså ingen markant forskjell når familieavhengighet tas med i betraktningen. Vi har også sett på analyser der vi har justert for fødselsår som en kontinuerlig variabel for å se om resultatene endres. Resultatene er ikke tatt med her, men de ble ikke så forskjellige fra resultatene i Tabell 4.1. Dette gjaldt alle dødsårsakene, både for kvinner og menn.

For alle dødsårsakene er det slik at den estimerte hazard ratioen minker når utdannelsesnivået øker. Dette gjelder for begge kjønn. For kvinner er det alkoholrelaterte årsaker og CHD som skiller seg ut. Det er en signifikant forskjell mellom de som har lav og høy utdannelse. Den estimerte hazard ratioen var 7.6 for kvinner som kun hadde fullført grunnskoleutdannelse for dødsårsaken CHD. Risikoen er altså tilnærmet åtte ganger høyere sammenlignet med kvinner som hadde universitets-og høyskoleutdannelse på høyere nivå. Fra Tabell A.4 i Kapittel 2 ser vi at CHD forekom hyppigst blant kvinner i alderen 50 til 59 år, og tilsvarende for menn. For alkoholrelaterte årsaker er denne raten lik 6.9 for kvinner med inntil niårig skole. Den samme raten er lik 1.3 for kvinner med høyskoleutdannelse. Begge konfidensintervallene for denne gruppen inneholder 1 og $\widehat{HR}$ ble lik 1.3. Dette tyder på at risikoen minker når utdannelsesnivået øker.

Når det gjelder menn, er også her alkoholrelaterte årsaker blant de dødsårsakene som gir størst forskjell mellom dødsrisiko for lavtutdannende sammenlignet med de som hadde høyskole- og universitetsutdannelse. Totalt var det 2 440 personer som døde på grunn av alkoholrelaterte årsaker, der 79% av disse var menn. Dødsfall på grunn av alkoholrelaterte årsaker utgjorde 4.8% av alle dødsårsakene. For denne dødsårsaken ble den estimerte hazard ration lik 5.9 for menn med kun grunnskoleutdannelse. Vi ser igjen at hazard ratioen øker når utdannelsesnivået avtar. Sammenlignet med de som hadde universitetsutdannelse ble $\widehat{HR}$ lik 3.9, 2.4 og 1.6 for henholdsvis yrkesskole, gymnas/artium og høyskole. Lungekreft hos menn er en annen dødsårsak som skiller seg ut blant de andre dødsårsakene. Her er dødsrisikoen er omtrent seks ganger høyere for menn med kun grunnskoleutdannelse i forhold til menn med universitetsutdannelse.

4.2 Resultater fra tilpasset Cox regresjonsmodell med stratifisering

Tabell 4.2 viser tilsvarende analyser som i Tabell 4.1, men her er det brukt Cox stratifiserte modell og det er stratifisert på søskenflokk. Også her er det separate analyser for begge kjønn. Datasettet besto av 721 805 personer uten "enebarn". Dette svarer til 292 300 søskenflokker. Som det er nevnt i Kapittel 3.3 er gruppene, eller i dette tilfellet søskenflokkene, uavhengige av hverandre. Resultatene her viser også at høyere utdannelse gir lavere risiko for dødelighet for alle dødsårsakene, men i forhold til Tabell 4.1 er likevel ikke denne reduksjonen like stor. For kvinner er det dødsårsakene lungekreft, CHD og alkoholrelaterte årsaker som gir størst forskjell mellom de analysene med og uten stratifisering

på søskenflokk. De største forskjellene mellom disse metodene gjelder for utdannelsesgruppen "7-9 år". Den estimerte hazard ratioen for denne gruppen i Tabell 4.1 ble 5.01, 7.60 og 6.91 for henholdsvis lungekreft, CHD og alkoholrelaterte årsaker. For den stratifiserte analysen ble de tilsvarende resultatene 3.45, 4.74 og 2.64. Dette viser at risikoen avtar når vi tar hensyn til familieavhengighet mellom søsken, og det gjelder spesielt alkoholrelaterte årsaker. For de andre utdannelsesnivåene er ikke forskjellene like store.

Tabell 4.2: *Cox regresjon med stratifisering på søskenflokk.*

		Kvinner			**Menn**		
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**
Total-dødelighet							
	7-9 år	2.06	0.09	1.72-2.47	2.37	0.06	2.10-2.69
	10-11 år	1.64	0.08	1.40-1.94	1.97	0.06	1.75-2.22
	12 år	1.41	0.09	1.18-1.68	1.52	0.06	1.36-1.71
	12-16 år	1.19	0.09	1.00-1.42	1.34	0.07	1.18-1.53
	>16 år	Ref.			Ref.		
Lungekreft							
	7-9 år	3.45	0.37	1.66-7.14	4.84	0.29	2.76-8.47
	10-11 år	2.58	0.36	1.28-5.17	4.26	0.28	2.46-7.39
	12 år	1.79	0.37	0.86-3.73	3.20	0.27	1.88-5.46
	12-16 år	1.14	0.36	0.56-2.35	3.46	0.30	1.94-6.17
	>16 år	Ref.			Ref.		
CVD							
	7-9 år	3.72	0.33	1.92-7.21	2.67	0.13	2.07-3.45
	10-11 år	2.75	0.32	1.48-5.12	2.08	0.13	1.63-2.66
	12 år	2.35	0.34	1.21-4.54	1.67	0.12	1.32-2.12
	12-16 år	1.41	0.35	0.72-2.80	1.24	0.14	0.94-1.64
	>16 år	Ref.			Ref.		
CHD							
	7-9 år	4.74	0.53	1.69-13.34	2.55	0.16	1.88-3.47
	10-11 år	3.36	0.50	1.26-8.97	1.98	0.15	1.48-2.66
	12 år	1.92	0.54	0.67-5.49	1.73	0.15	1.29-2.30
	12-16 år	1.38	0.59	0.44-4.35	1.14	0.17	0.82-1.59
	>16 år	Ref.			Ref.		
Alkoholrelaterte årsaker							
	7-9 år	2.64	0.72	0.65-10.77	3.27	0.28	1.87-5.72
	10-11 år	1.11	0.67	0.30-4.11	2.40	0.27	1.41-4.08
	12 år	1.04	0.66	0.28-3.81	1.25	0.27	0.73-2.15
	12-16 år	1.04	0.71	0.26-4.16	0.92	0.31	0.50-1.71
	>16 år	Ref.			Ref.		
Ulykker							
	7-9 år	2.46	0.37	1.19-5.07	2.35	0.21	1.57-3.51
	10-11 år	1.69	0.34	0.86-3.31	2.00	0.19	1.37-2.92
	12 år	1.68	0.37	0.81-3.47	1.62	0.19	1.12-2.35
	12-16 år	1.06	0.39	0.50-2.28	1.13	0.21	0.71-1.79
	>16 år	Ref.			Ref.		

For menn er det de samme dødsårsakene som for kvinner når det gjelder forskjellen mellom hazard ratioene for Cox regresjon med og uten stratifisering i tabellene 4.1 og 4.2. Den største forskjellen gjelder for alkoholrelaterte årsaker. Her ble $\widehat{HR}$ for menn med kun grunnskoleutdannelse lik 3.27 og 5.95 for henholdsvis Cox regresjon med og uten stratifisering på søskenflokk. Konfidensintervallene i Tabell 4.2 er bredere enn de tilsvarende konfidensintervallene i Tabell 4.1. Det er flere årsaker til at det er slik. Flere av personene i datasettet faller ut fra denne analysen. Ved stratifisering på søskenflokk, blir hver søskenflokk et stratum. I dette datasettet kan hver søskenflokk bestå av maksimum 4 søsken, og siden vi har egne analyser for kvinner og menn, må det være søsken av samme kjønn. Det blir da små grupper og usikkerheten blir større. I tillegg vil ikke et stratum som består av én person, altså "enebarn" bidra i analysene. De personene som ikke dør i løpet av studien vil heller ikke bidra. Hvis en søskenflokk består av to personer og levetiden til den eldste er større enn sensureringstiden til den yngste, vil dette føre til at også slike søskenflokker ikke vil bidra i de søskenstratifiserte analysene. Dermed vil det være færre personer i denne analysen sammenlignet med forrige seksjon, og usikkerheten er større.

4.3 *Variance of random effect* og tetthetsplott for gamma- og log-normal frailty

En måte å analysere multivariate data på er å bruke frailty-modeller. Disse modellene tar hensyn til familieavhengigheten på en annen måten enn de modellene vi har brukt hittil. I slike modeller antar man at alle personene i samme gruppe har samme frailty. Disse modellene er beskrevet i Kapittel 3. I dette tilfellet gjelder det hver søskenflokk. Det finnes flere fordelinger for frailty-variabelen. I denne oppgaven bruker vi at frailty-variabelen Z er gamma- og log-normalfordelt. Resultatene fra disse analysene er vist videre i Seksjon 4.4, Seksjon 4.5 og Tillegg C.

I Tabell 4.3 er θ_G og θ_{LN}, som er variansen til henholdsvis en gamma-fordelt frailty- og log-normal frailty-variabel, listet opp for de ulike dødsårsakene. Dette er *variance of random effect*. En høy verdi av θ indikerer mer avhengighet innad i gruppene. Dette betyr at familieavhengigheten er viktigere jo høyere denne variansen er. Figurene 4.1, 4.2, 4.3 og 4.4 og illustrerer endringen av tettheten for de to nevnte fordelingene når θ_G og θ_{LN} endres for de ulike dødsårsakene. Her er tetthetene, som vises i Formel (3.28) og Formel (3.33) i Kapittel 3, plottet for z-verdier mellom 0 og 5. Plottene for totaldødelighet, CHD, lungekreft og alkoholrelaterte årsaker er tatt med i denne seksjonen, mens for CVD og ulykker er disse tatt med i Tillegg C.1 (Figur C.1 og Figur C.2). I begge fordelingene er det parametrisert slik at forventningen blir lik 1. I Formel (3.29) er forventningen lik 1 og variansen lik $\theta = \theta_G$. For log-normalfordelingen er variansen vist i Formel (4.1). Vi bruker θ_{LN} fra `R`-utskriften og setter inn i formelen for variansen. Videre setter vi forventningen lik 1, og løser ligningen slik at vi får et uttrykk for μ med σ. Ved å bruke uttrykkene for forventningen og variansen ved log-normal frailty, kan vi uttrykke σ og μ slik:

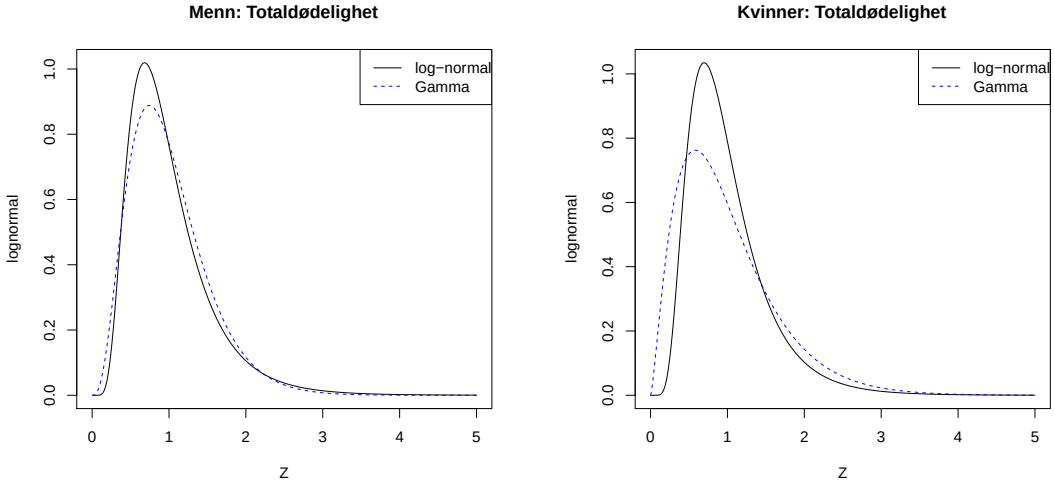
$$e^{\sigma^2} - 1 = \theta_{LN} \iff \sigma = \sqrt{\log(\theta_{LN} + 1)} \tag{4.1}$$

$$e^{\mu + \frac{1}{2}\sigma^2} = 1 \iff \mu = -\frac{1}{2}\sigma^2. \tag{4.2}$$

Fra Formel (3.29) med gamma frailty ser vi at når $\theta_G = 1$, får vi eksponensial-fordelingen med parameter lik 1 når vi antar at frailty-variabelen er gamma-fordelt. Når θ blir mindre enn 1, blir det en topp og $f_Z(z)$ går mot normalfordelingen, mens for θ-verdier større enn 1 går $f_Z(z)$ mot uendelig. Som Tabell 4.3 viser, er det stor forskjell mellom θ_G og θ_{LN} både for kvinner og menn. Dette vises også gjennom tetthetsplottene. Den minste forskjellen er for totaldødelighet for begge kjønn. De største forskjellene gjelder for lungekreft, CHD og alkoholrelaterte årsaker. For kvinner er det $\theta_G = 4.97$ og $\theta_{LN} = 1.06$, begge for CHD, som er de største verdiene. For menn er de største effektene for alkoholrelaterte årsaker og ulykker, der variance of random effect ble henholdsvis $\theta_G = 3.96$ og $\theta_{LN} = 0.63$. Resultatene fra neste seksjon viser at det ikke er så forskjell mellom gamma- og log-normal frailty. Dette stemmer ikke overrens med Tabell 4.3. I denne tabellen er det svært stor forskjell mellom θ_G og θ_{LN}. Denne forskjellen klarer vi ikke å forklare.

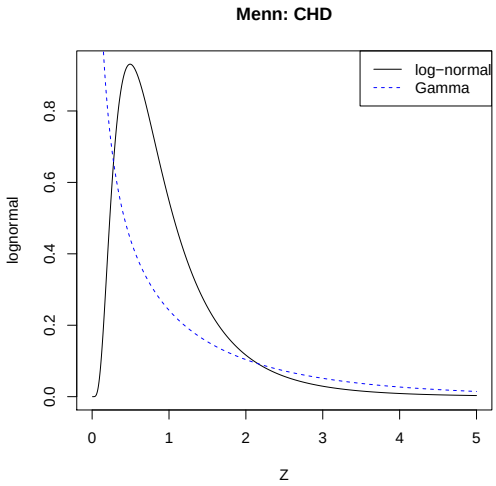
Tabell 4.3: *Variance of random effect for gamma- og log-normal frailty.*

	θ_G: Gamma		θ_{LN}: Log-normal	
Dødsårsak	**Kvinner**	**Menn**	**Kvinner**	**Menn**
Totaldødelighet	0.42	0.26	0.27	0.29
Lungekreft	3.07	1.60	0.89	0.61
CVD	2.00	1.29	0.63	0.55
CHD	4.97	2.00	1.06	0.60
Alkoholrelaterte årsaker	0.50	3.96	0.04	0.49
Ulykker	0.50	2.00	0.43	0.63

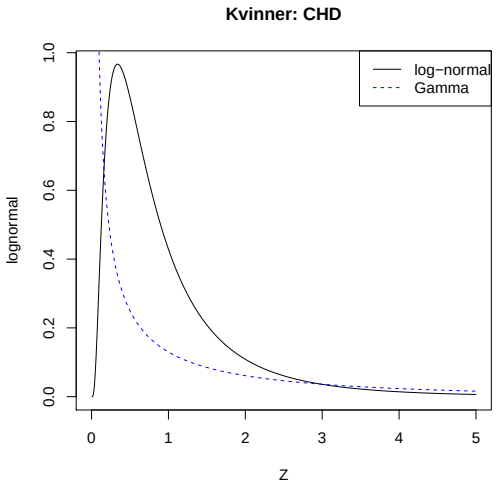


(a) (b)

Figur 4.1: (a) $\theta_G = 0.26$, $\theta_{LN} = 0.29$ (b) $\theta_G = 0.42$, $\theta_{LN} = 0.27$

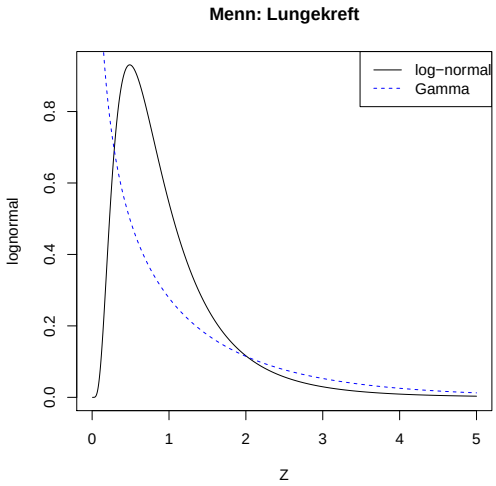


(a)

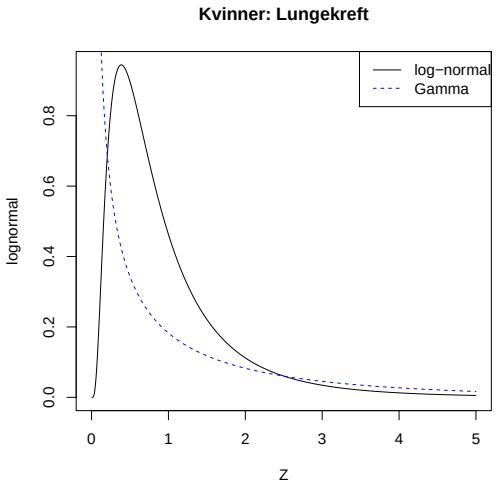


(b)

Figur 4.2: (a) $\theta_G = 2.00$, $\theta_{LN} = 0.60$ (b) $\theta_G = 4.97$, $1\theta_{LN} = 1.06$

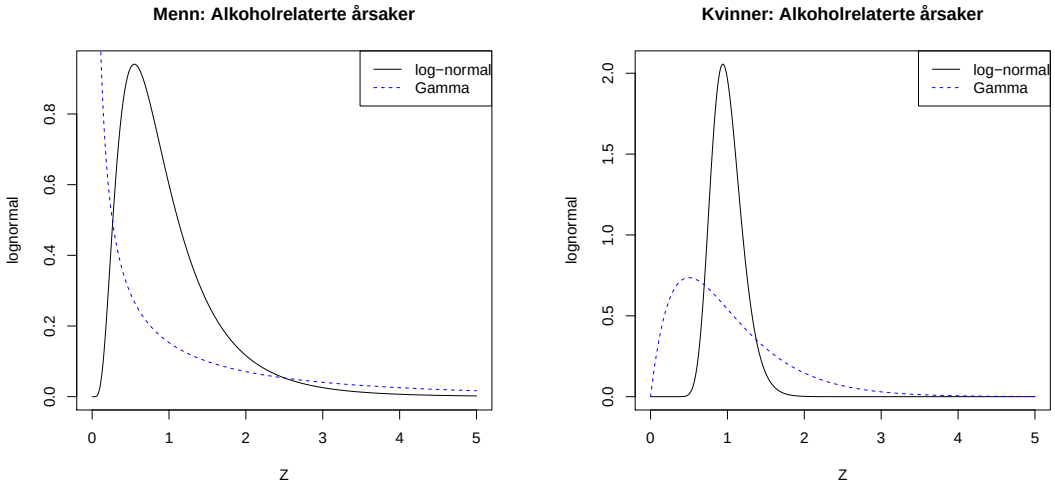


(a)



(b)

Figur 4.3: (a) $\theta_G = 1.60$, $\theta_{LN} = 0.61$ (b) $\theta_G = 3.07$, $\theta_{LN} = 0.89$



(a) (b)

Figur 4.4: (a) $\theta_G = 3.96$, $\theta_{LN} = 0.49$ (b) $\theta_G = 0.50$, $\theta_{LN} = 0.04$

4.4 Sammenligning av Cox regresjonsmodeller og frailty-modeller for totaldødelighet

I Tabell 4.4 sammenlignes flere metoder for totaldødelighet. Igjen er utdannelse delt inn i fem kategorier og det er utdannelsesgruppen ">16 år" (universitet) som er referansegruppen. I denne tabellen er det den estimerte verdien av $\boldsymbol{\beta}$ og standardfeilen som sammenlignes for de ulike metodene for begge kjønn. I den første metoden ser vi bort fra at datasettet består av søskenflokker og antar at det ikke er noe avhengighet mellom de 934 548 personene, mens for de tre siste metodene antar vi at det er avhengighet mellom de 505 043 gruppene med familier. Denne familieavhengigheten blir tatt hensyn til på ulike måter for de metodene som er anvendt her. I Tabell 4.1 så vi at det ikke var en markant forskjell i resultatene for Cox regresjon når vi antar uavhengighet eller tilpasser en Cox regresjonsmodell med en sandwich-estimator. I tabellene videre er det derfor kun tatt med resultater for Cox regresjon uten å ta hensyn til at dataene kan være korrelerte. Fra Tabell 4.4 ser vi at de estimerte verdiene av regresjonskoeffisientene minker når utdannelsesnivået øker. De er høyere for menn enn kvinner, mens usikkerheten ble høyere for kvinner. Dette gjelder alle metodene og alle utdannelsesnivåene. Det ser ut som alle de fire metodene gir tilnærmet de samme resultatene, men standardfeilene ble størst for den stratifiserte modellen på grunn av få personer i hvert stratum, mens de er like for de andre metodene. Ved å sammenligne Cox regresjonsmodell med og uten stratifisering på søskenflokk, ser vi at resultatene fra disse regresjonsmetodene ikke avviker så mye for kvinner. For menn er det litt annerledes. De estimerte hazard ratioene er lavere for stratifisert Cox-modell enn de andre metodene. Dette gjelder de tre første utdannelsesnivåene. P-verdien for kvinner med høyskoleutdannelse ble 5% i den stratifiserte analysen. De resterende p-verdiene ble tilnærmet 0. Videre i dette kapittelet skal vi teste om denne forskjellen skyldes tilfeldigheter eller ikke. For å gjøre dette skal vi brukes bootstrapping. Denne metoden er forklart

Tabell 4.4: *Sammenligning av metoder for totaldødelighet.*

		Kvinner		Menn	
Metode	**Utdannelse**	$\hat{\beta}$	**se**($\hat{\beta}$)	$\hat{\beta}$	**se**($\hat{\beta}$)
Cox					
	7-9 år	0.78	0.03	1.05	0.02
	10-11 år	0.41	0.03	0.80	0.02
	12 år	0.32	0.04	0.51	0.02
	12-16 år	0.13	0.04	0.28	0.02
	>16 år	Ref.		Ref.	
Stratifisert Cox (Stratifisert på søskenflokk)					
	7-9 år	0.72	0.09	0.86	0.06
	10-11 år	0.50	0.08	0.68	0.06
	12 år	0.34	0.09	0.42	0.06
	12-16 år	0.17	0.09	0.30	0.07
	>16 år	Ref.		Ref.	
Frailty: Gamma					
	7-9 år	0.79	0.03	1.06	0.02
	10-11 år	0.41	0.03	0.81	0.02
	12 år	0.32	0.04	0.51	0.02
	12-16 år	0.14	0.04	0.28	0.03
	>16 år	Ref.		Ref.	
Frailty: Log-normal					
	7-9 år	0.79	0.03	1.06	0.02
	10-11 år	0.41	0.03	0.81	0.02
	12 år	0.32	0.04	0.51	0.02
	12-16 år	0.14	0.04	0.28	0.03
	>16 år	Ref.		Ref.	

i Kapittel 3.6. For totaldødeligheten er det ikke noen forskjell i resultatene for gamma frailty og log-normal frailty. Som nevnt over, blir også grupper med en person tatt med i disse analysene. Vi har også sett på analyser der familier med kun ett barn i hver gruppe er tatt ut av datasettet. Dette er vist i Tabell C.3 (Tillegg C.3). Her har vi brukt at frailty-variabelen er gamma-fordelt. I tabellene C.1 og C.2 sammenligner hazard ratioene og 95% konfidensintervaller for alle dødsårsakene for gamma- og log-normal frailty. Fra disse tabellene ser vi at resultatene ble svært like. Siden det ikke var så store avvik mellom resultatene for gamma og log-normal, har vi kun sett på analyser for en gamma-fordelt frailty-variabel i Tabell C.3 for å se om det gir noen effekt ved å ikke ta med "enebarn" i analysene. Usikkerheten ble noe større da "enebarn" ble ekskludert fra analysene. Dette gjaldt for begge kjønn. For kvinner ble det ikke så stor endring i resultatene. De største avvikene var for de to første utdannelsesnivåene for alkoholrelaterte årsaker. For menn var det tilsvarende for CVD.

4.5 Sammenligning av Cox regresjonsmodeller og frailty-modeller for CHD

For hjerteinfarkt ble de fleste p-verdiene avrundet til 0 for alle utdannelsesnivåene for menn, bortsett fra høyskole ("12-16 år") i den stratifiserte analysen. Da ble p-verdien lik 0.44. For kvinner ble mange av kovariatene signifikante. De p-verdiene som ikke ble 0 var for høyskoleutdannelse for alle metodene. P-verdiene ble henholdsvis 0.38, 0.58, 0.37 og 0.38 for gruppen "12-16 år" i den rekkefølgen de er skrevet i Tabell 4.5. Fra tabellen ser vi blant annet at resultatene fra Cox, gamma frailty og log-normal frailty er like for begge kjønn. Når vi stratifiserer på søskenflokk, skiller resultatene seg ut blant de andre metodene for denne dødsårsaken. For den stratifiserte analysen er β-verdiene noe lavere, mens standardavvikene er høyere. For eksempel ble den estimerte hazard ratioen lik $e^{1.56} = 4.8$ for gruppen "7-9 år" i den stratifiserte analysen, mens for de resterende metodene ble estimert hazard ratio $e^{2.03} = 7.6$. Dermed er dødsrisikoen lavere for denne utdannelsesgruppen når vi tar hensyn til avhengighet mellom søsken på denne måten. Det er interessant at risikoen er den samme for frailty-resultatene der vi tar hensyn til familieavhengigheten som for Cox regresjon der vi antar at alle personene er uavhengige av hverandre.

For alkoholrelaterte årsaker og lungekreft har vi ikke tatt med resultatene her. De er oppsummert i Tabell C.4 (Tillegg C.4) og Tabell C.5 (Tillegg C.5). Fra Tabell C.4 ser vi at også her er det den andre metoden som gir veldig ulike resultater enn de andre metodene. Usikkerheten er mye større og de estimerte koeffisientene er lavere. På nivå 5% ble kun grunnskole signifikant. P-verdiene for kvinner i den stratifiserte analysen ble 0.18, 0.87, 0.95 og 0.95 for henholdsvis grunnskole, yrkesskole, gymnas/artium og høyskole.

Lungekreft er den mest utbredte typen kreft i Norge[1], og noen av årsakene til denne krefttypen er nevnt i Kapittel 2 Fra Tabell C.5 legger vi merke til at risikoen for å dø av lungekreft minker når utdannelsesnivået øker, både for kvinner og menn. Sammenlignet med de som har universitetsutdannelse, er det de med høyskoleutdannelse som har minst

[1] http://www.ssb.no/vis/samfunnsspeilet/utg/201102/15/art-2011-05-02-01.html

Tabell 4.5: *Sammenligning av metoder for CHD.*

		Kvinner		**Menn**	
Metode	**Utdannelse**	$\hat{\beta}$	**se**($\hat{\beta}$)	$\hat{\beta}$	**se**($\hat{\beta}$)
Cox					
	7-9 år	2.03	0.20	1.33	0.06
	10-11 år	1.27	0.20	0.98	0.06
	12 år	0.99	0.21	0.68	0.06
	12-16 år	1.24	0.25	0.36	0.07
	>16 år	Ref.		Ref.	
Stratifisert Cox (Stratifisert på søskenflokk)					
	7-9 år	1.56	0.53	0.93	0.16
	10-11 år	1.21	0.50	0.68	0.15
	12 år	0.76	0.54	0.55	0.15
	12-16 år	0.32	0.59	0.13	0.17
	>16 år	Ref.		Ref.	
Frailty: Gamma					
	7-9 år	2.04	0.20	1.35	0.06
	10-11 år	1.27	0.20	0.99	0.06
	12 år	1.00	0.22	0.69	0.06
	12-16 år	0.22	0.25	0.37	0.07
	>16 år	Ref.		Ref.	
Frailty: Log-normal					
	7-9 år	2.03	0.20	1.34	0.05
	10-11 år	1.27	0.20	0.98	0.06
	12 år	0.99	0.21	0.68	0.06
	12-16 år	0.22	0.25	0.36	0.07
	>16 år	Ref.		Ref.	

risiko for å dø av lungekreft. For eksempel ble den estimerte hazard-ratioen lik $e^{1.61} = 5.00$ og $e^{0.13} = 1.14$ for kvinner som hadde fullført henholdsvis grunnskole og høyskole sammenlignet med kvinner med universitetsutdannelse. Ved en sammenligning av metodene ser vi at resultatene fra Cox regresjon uten stratifisering og frailty blir tilnærmet like. Som for CHD, er det større avvik mellom resultatene fra den stratifiserte analysen og de tre andre metodene. I tillegg er dødsrisikoen høyere for menn enn kvinner. Dette gjelder alle utdannelsesnivåene.

4.6 Sammenligning av resultater fra tilpasset Cox regresjonsmodell for "enebarn" og søsken

I denne seksjonen sammenligner vi analyser for de personene som er "enebarn" og søsken. Som nevnt i innledningen, er grupper på en person ikke nødvendigvis enebarn. De kan ha søsken som er født før 1940 eller etter 1959, og som da ikke er med i studien. Resultatene er oppsummert i Tabell 4.6 for lungekreft og alkoholrelaterte årsaker, og i Tabell D.1 (Tillegg D.1) for de resterende dødsårsakene. For "enebarn", som utgjorde 22.8% (212 743) av datasettet, er det tilpasset en Cox regresjonsmodell uten stratifisering i R der vi antar at alle personene er uavhengige av hverandre. Når det gjelder søsken, har vi brukt Cox regresjon der også, men i tillegg har vi tatt hensyn til familieavhengigheten blant søsken. Her har vi, som i de innledende analysene, brukt en sandwich-estimator. De estimerte hazard ratioene blir de samme som når vi antar uavhengighet, men konfidensintervallene blir annerledes. I dette datasettet er det 292 300 søskenflokker. Disse utgjør 721 805 personer. Resultatene for analyser der vi antar at alle de 721 805 personene er uavhengige av hverandre er ikke tatt med her, men for disse analysene ble forskjellen mellom 95% konfidensintervaller for HR fra Cox regresjonsmodell med og uten sandwich-estimatoren svært liten.

I Tabell D.2 i Tillegg D.1 har vi testet om det er en signifikant forskjell mellom resultatene for "enebarn" og søsken. Dette har vi gjort ved å finne $z = \frac{\hat{\beta}_{\text{"Enebarn"}} - \hat{\beta}_{\text{Søsken}}}{\sqrt{\text{se}_{\text{"Enebarn"}} - \text{se}_{\text{Søsken}}}}$ og p-verdi for alle utdannelsesnivåene der nullhypotesen er $\hat{\beta}_{\text{"Enebarn"}} - \hat{\beta}_{\text{Søsken}} = 0$. Denne tabellen viser at De tre første utdannelsesnivåene for totaldødelighet ga forkastning på nivå 5% for begge kjønn. For menn var det CVD og alkoholrelaterte årsaker som ble signifikante for alle utdannelsesnivåene. For lungekreft gjaldt de samme for de tre første gruppene. For kvinner var det kun gruppen "10-11 år" og "7-9 år" for henholdsvis alkoholrelaterte årsaker og ulykker som ga forkastning på nivå 5%.

En sammenligning av resultatene for totaldødeligheten viser at det er forskjell mellom estimert hazard ratio for søsken og "enebarn" for begge kjønn. Estimatene er høyere og konfidensintervallene for HR er bredere for "enebarn" enn søsken. Det var resultatene for lungekreft og alkoholrelaterte årsaker som skilte seg mest ut sammenlignet med de andre dødsårsakene. For CVD, CHD og ulykker er resultatene oppsummert i Tabell 4.6 i Tillegg D.1. Når vi sammenligner resultatene for kvinner i tabellene 4.6 og D.1, ser vi at $\widehat{HR}$ er høyere for "enebarn" enn søsken i de fleste tilfellene. For alkoholrelaterte årsaker er den største forskjellen for de to første utdannelsesnivåene, der estimert hazard ratio ble

7.67 og 5.20 for henholdsvis gruppene "7-9 år" og "10-11 år" for "enebarn". De tilsvarende verdiene de gruppene for søsken ble 6.71 og 2.68.

For menn er det også alkoholrelaterte årsaker som gir de største differansene mellom hazard-ratioene. Dette gjelder utdannelsesnivåene grunnskole og yrkesskole. Her er dødsrisikoen ca. ni ganger høyere ($\widehat{HR} = 8.76$) for menn som er "enebarn" og som kun har grunnskoleutdannelse sammenlignet med menn som også er "enebarn", men som har universitetsutdannelse. Den tilsvarende risikoen for menn som har søsken er 5.34. Totalt sett er det slik at utdannelseseffekten er sterkere blant "enebarn" enn søsken. Dette gjelder spesielt dødsårsakene lungekreft, CHD og alkoholrelaterte årsaker. Det er stor forskjell mellom dødsrisikoen for de som har lav utdannelse sammenlignet med de personene som har høyskole-og universitetsutdannelse. Dødsrisikoen minker når utdannelsesnivået øker. For søsken er disse resultatene ikke så ulike som de i Tabell 4.1, der "enebarn" ble inkludert i analysene.

Tabell 4.6: *Cox regresjon uten stratifisering for "enebarn" og søsken. For søsken er det tatt hensyn til familieavhengighet.*

		Kvinner				**Menn**			
		"Enebarn"		**Søsken**		**"Enebarn"**		**Søsken**	
Dødsårsak	**Utd.**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**
Total-dødelighet									
	7-9 år	2.40	2.11-2.73	2.14	1.99-2.29	3.07	2.80-3.35	2.81	2.67-2.96
	10-11 år	1.70	1.50-1.93	1.44	1.35-1.55	2.44	2.23-2.66	2.16	2.05-2.27
	12 år	1.55	1.35-1.79	1.31	1.21-1.42	1.76	1.61-1.93	1.63	1.55-1.71
	12-16 år	1.21	1.04-1.40	1.13	1.03-1.22	1.39	1.25-1.55	1.30	1.22-1.38
	>16 år	Ref.		Ref.		Ref.		Ref.	
Lungekreft									
	7-9 år	5.39	3.23-8.99	4.91	3.70-6.51	4.34	3.09-6.09	6.06	4.84-7.59
	10-11 år	2.24	1.34-3.75	2.66	2.00-3.52	3.39	2.41-4.77	4.04	3.21-5.08
	12 år	2.25	1.28-3.93	2.06	1.50-2.82	2.49	1.76-3.52	3.18	2.53-4.00
	12-16 år	1.01	0.52-1.95	1.19	0.84-1.69	1.95	1.30-2.91	2.27	1.74-2.97
	>16 år	Ref.		Ref.		Ref.		Ref.	
Alkohol-relaterte årsaker	7-9 år	7.67	2.78-21.17	6.71	3.73-12.08	8.76	5.33-14.41	5.34	4.19-6.79
	10-11 år	5.20	1.90-14.21	2.68	1.48-4.85	6.09	3.69-10.04	3.47	2.71-4.44
	12 år	2.73	0.91-8.18	2.22	1.15-4.28	4.16	2.51-6.89	2.00	1.55-2.57
	12-16 år	1.05	0.28-3.92	1.37	0.66-2.82	2.49	1.39-4.45	1.37	1.00-1.88
	>16 år	Ref.		Ref.		Ref.		Ref.	

Det er også interessant å sammenligne resultatene for søsken i Tabell 4.6 med søsken-stratifiserte analyser i Tabell 4.2. Som beskrevet i Seksjon 4.2, blir usikkerheten større ved stratifisering. For kvinner og totaldødelighet er hazard ratioen høyere for alle utdannelsesnivåene bortsett fra grunnskole, men disse forskjellene er ikke så store. For menn er det kun $\widehat{HR}$ for høyskole som er høyere, men som for kvinner, er konfidensintervallene bredere. Når det gjelder lungekreft og alkoholrelaterte årsaker, er forskjellene større enn

for totaldødelighet. For eksempel er hazard ratioen for kvinner med kun grunnskoleutdannelse lik 3.45 i den stratifiserte analysen, mens det tilsvarende tallet i Tabell 4.6 for søsken er 4.91. De største forskjellene mellom Tabell 4.2 og Tabell 4.6 finner vi for alkoholrelaterte årsaker. Usikkerheten er større for denne dødsårsaken, fordi det var få dødsfall på grunn av alkoholrelaterte årsaker. Her er resultatene for søsken i Tabell 4.6 sammenlignet med resultatene fra den søsken-stratifiserte analysen høyere for alle utdannelsesnivåene for begge kjønn. Dødsrisikoen er mindre når vi tar hensyntar søskenavhengigheten ved stratifisering enn når vi bruker en sandwich-estimator.

Analysene i denne seksjonen er de samme som i artikkelen[1]. Som nevnt i innledningen, var det kun søskenflokker som ble tatt med. Dersom man sammenligner resultatene fra Tabell 4.6 med Tabell 3 på side 7 i artikkelen (*Cohort*), er resultatene for totaldødelighet tilnærmet like. Dette gjelder både kvinner og menn. For lungekreft er mellom resultatene her og de i artikkelen forskjellen noe større. Blant annet er dødsrisikoen for alle utdannelsesgruppene for lungekreft og alkoholrelaterte årsaker noe høyere for kvinner i Tabell 4.6 sammenlignet med tabellen i artikkelen. De største forskjellene gjelder den første utdannelsesgruppen. Når det gjelder resultatene for menn, er dødsrisikoen for alkoholrelaterte årsaker høyere i Tabell 4.6 enn i artikkelen. For eksempel ble den estimerte hazard ratioen for menn med kun grunnskoleutdannelse lik 7.67 og 5.47 for henholdsvis denne tabellen og Tabell 3 i artikkelen. Avvikene kan skyldes at datasettet som brukes i denne oppgaven er noe annerledes enn det som ble brukt i studien. Det var 337 627 søskenflokker (871 367 personer) i studien, mens det her er 292 300 søskenflokker(721 805 personer). Dessuten ble søskenflokker som besto av flere enn fire personer tatt med, i motsetning til her, hvor alle søskenflokker med fem søsken eller flere ble fjernet fra datasettet.

Kapittel 5

Resultater fra bootstrapping

I dette kapittelet har vi brukt bootstrap-metoden som er beskrevet i Kapittel 3.6. Hensikten er å teste om det er systematiske forskjeller mellom resultatene fra Cox regresjonsmodell med og uten stratifisering. Hvis det er forskjell, kan det tyde på at søskenavhengigheten er viktig for dødsrisikoen. For å se om endringen av antall bootstrap-utvalg gir utslag, er det blitt brukt både 100 og 500 bootstrap-utvalg. Som skrevet i Seksjon 3.6, kan vi også sammenligne disse resultatene med de vi har fått tidligere for å se om estimatoren er forventningsskjev. Her vises resultatene for $B = 500$ bootstrap-utvalg. Som i forrige kapittel, konsentrerer vi oss om totaldødelighet og CHD, alkoholrelaterte årsaker og lungekreft på grunn av at det var disse dødsårsakene som ga de største hazard ratioene. For hver dødsårsak er det først en tabell som viser gjennomsnittet av bootstrap-parameterne, $\frac{1}{B}\sum_{i=1}^{B}\hat{\beta}_i^*$ i tillegg til standardfeilen separat for kvinner og menn. Dette er vist i Tabell 5.1 og Tabell 5.3 for henholdsvis totaldødelighet og CHD. I tabellene E.1 og E.3 i Tillegg E er de tilsvarende resultatene for 100 bootstrap-utvalg angitt. Her er også resultatene for alkoholrelaterte årsaker og lungekreft for både 100 og 500 bootstrap-utvalg vist.

I tabellene 5.2, E.6 og E.8 har vi estimert variansen

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}^* - \hat{\boldsymbol{\beta}}_{\mathbf{2}}^*\right) = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}^*\right) + \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{2}}^*\right) - 2\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}^*, \hat{\boldsymbol{\beta}}_{\mathbf{2}}^*\right). \tag{5.1}$$

Dette er variansen mellom boostrap-estimatene for forskjellen mellom Cox regresjonsmodell og Cox stratifisert regresjonsmodell. Vi definerer $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^*$ og $\hat{\boldsymbol{\beta}}_{\mathbf{2}}^*$ til å være to vektorer med lengde B som består av de estimerte koeffisientene fra henholdsvis Cox regresjon og stratifisert Cox regresjon. Som tidligere nevnt blir alle personene tatt med i analysen, mens når det stratifiseres på søskenflokk faller "enebarn" bort fra analysen. Tabell 5.2 viser variansen for totaldødelighet og CHD, mens tabellene E.6 og E.8 viser det samme for henholdsvis alkoholrelaterte dødsårsaker og lungekreft.

Figur 5.1 - Figur 5.4 viser de estimerte for bootstrap-parameterne for totaldødelighet og CHD for de ulike utdannelsesnivåene. De estimerte $\boldsymbol{\beta}$-verdiene fra Cox regresjon er plottet langs x-aksen, mens $\hat{\boldsymbol{\beta}}$-verdiene fra den stratifiserte analysen er plottet langs y-aksen. De tilsvarende figurene for alkoholrelaterte årsaker og lungekreft er vist i henholdsvis Tillegg E.3 og Tillegg E.4. I tillegg har vi beregnet korrelasjonen i Formel (5.2), som er

korrelasjonen mellom $\hat{\boldsymbol{\beta}}$-verdiene fra Cox regresjon med og uten stratifisering på søskenflokk. Dette er gjort for $B = 500$ bootstrap-utvalg. Jo høyere denne korrelasjonen er, desto sterkere er sammenhengen mellom estimatene. Figurene for $B = 100$ bootstrap-utvalg er i Tillegg E. I siste del av dette kapittelet er det en tabell som angir persentil-konfidensintervaller. I dette tilfellet er det persentilmetoden som er brukt. Fremgangsmåten er beskrevet i Seksjon 3.6.2.

$$\text{Corr}\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{*}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{*}\right) = \frac{\text{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{*}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{*}\right)}{\sqrt{\text{Var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{*}\right)\text{Var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{2}}^{*}\right)}}, \tag{5.2}$$

5.1 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering for totaldødelighet

For totaldødelighet ser parameterne ut til å være nokså like for $B = 100$ og $B = 500$ bootstrap-utvalg for Cox regresjon, både for kvinner og menn. Det samme er tilfellet når vi tilpasser en stratifisert Cox-modell. Resultatene i denne tabellen stemmer godt overens med estimatene i Tabell 4.4. Som nevnt over, viser Figur 5.1 og Figur 5.2 spredningsplott med korrelasjon som vist i Formel (5.2). Generelt viser plottene at det er større variasjon mellom punktene for kvinner enn menn. For 100 bootstrap viser plottene at korrelasjonen er høyere for kvinner enn menn for alle utdannelsesnivåene, bortsett fra grunnskole i Figur E.1a. Der ble korrelasjonen lik 0.45 og 0.43 for henholdsvis kvinner og menn. For bootstrap-utvalg lik 500, er korrelasjonen tilnærmet lik for begge kjønn for grunnskole og yrkesskole og igjen lavere for menn for de to siste utdannelsesgruppene.

Tabell 5.1: *Sammenligning av Cox-modell med og uten stratifisering for totaldødelighet.*

		500 bootstrap-utvalg			
		Kvinner		**Menn**	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^{*}$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^{*}$	$s_{\hat{\beta}}$
Cox					
	7-9 år	0.78	0.03	1.05	0.02
	10-11 år	0.41	0.03	0.80	0.02
	12 år	0.31	0.04	0.51	0.02
	12-16 år	0.13	0.04	0.28	0.03
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	0.72	0.10	0.86	0.07
	10-11 år	0.50	0.09	0.67	0.06
	12 år	0.34	0.10	0.42	0.06
	12-16 år	0.17	0.10	0.29	0.06
	>16 år	Ref.		Ref.	

For å se om det er en signifikant forskjell mellom resultatene fra analysene når vi tilpasser Cox regresjon med resultatene fra den stratifiserte analysen, finner vi variansen i Formel (5.1). Resultatene er vist i Tabell 5.2. Varians-estimatene er relativt små for begge kjønn, noe høyere for kvinner enn menn. Her endres ikke resultatene så mye når antall bootstrap-utvalg økes fra 100 til 500.

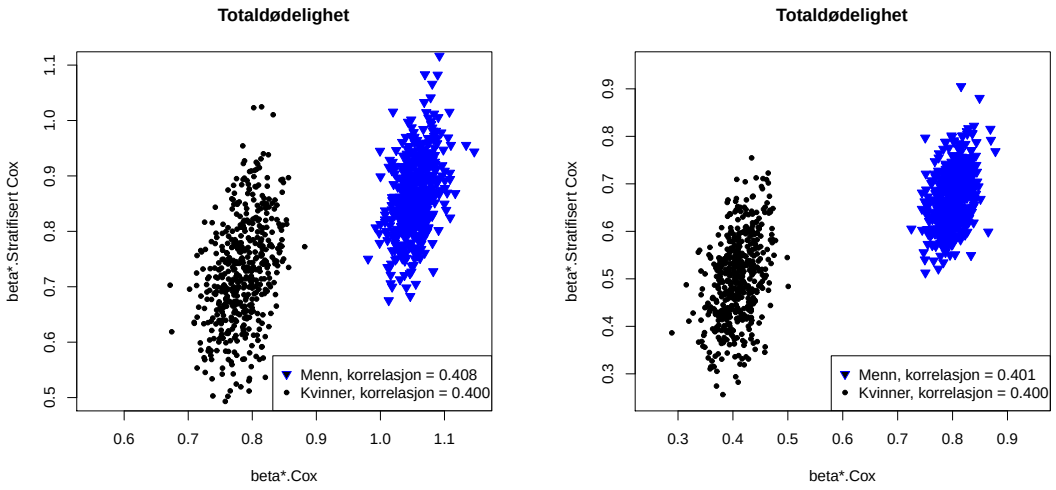
Tabell 5.2: *Variansen for totaldødelighet og CHD.*

		500 bootstrap-utvalg	
Dødsårsak	**Utdannelse**	**Kvinner**	**Menn**
Totaldødelighet			
	7-9 år	0.008	0.004
	10-11 år	0.007	0.003
	12 år	0.007	0.003
	12-16 år	0.008	0.003
	>16 år	Ref.	Ref.
CHD			
	7-9 år	2.858	0.023
	10-11 år	2.826	0.020
	12 år	2.869	0.021
	12-16 år	2.836	0.024
	>16 år	Ref.	Ref.

5.2 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering for CHD

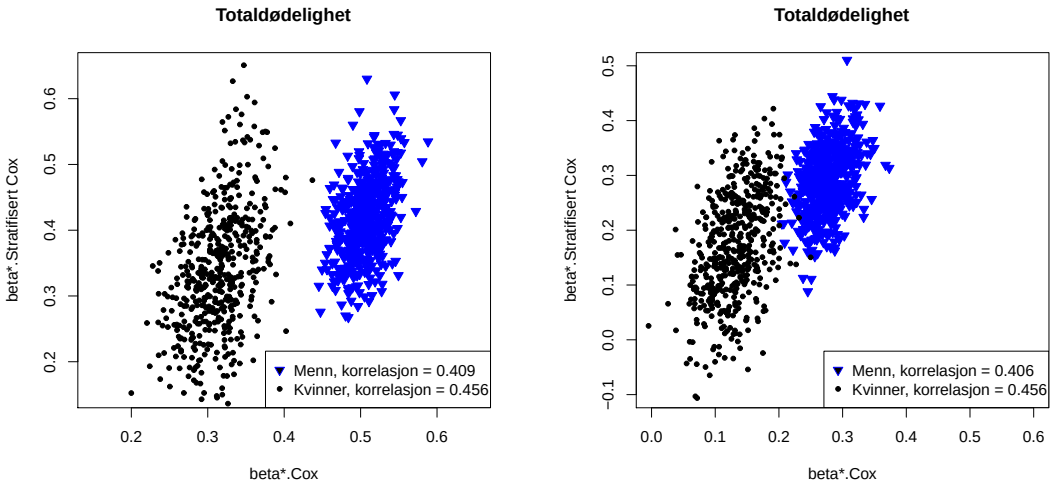
For CHD var det flest menn med utelukkende grunnskoleutdannelse som døde i løpet av studien. Dette er vist i Tabell 2.6 i Kapittel 2. Ved å sammenligne Tabell 5.3 og Tabell E.3 i Tillegg E, ser det ikke ut som det er så stor forskjell mellom resultatene for B =100 og B=500 bootstrap-utvalg. For den stratifiserte analysen for kvinner ble derimot standardfeilen større sammenlignet med Cox regresjon. Dette kan man også se fra figurene 5.3, 5.4, E.3 og E.10.

Som vi ser fra Figur 5.3 og Figur 5.4, er korrelasjonen mellom estimatene høyere for menn enn kvinner, mens variasjonen er mindre. Fra Tabell 5.2 ser vi at variansen fra Formel (5.1) er mye høyere for kvinner enn menn for denne dødsårsaken. Den er tilnærmet 2.9 og 0.02 for alle utdannelsesnivåene for henholdsvis kvinner og menn. For alkoholrelaterte årsaker og lungekreft er resultatene angitt i henholdsvis Tillegg E.3 og Tillegg E.4.



(a) (b)

Figur 5.1: (a) *7-9 års utdannelse: 500 bootstrap* (b) *10-11 års utdannelse: 500 bootstrap*

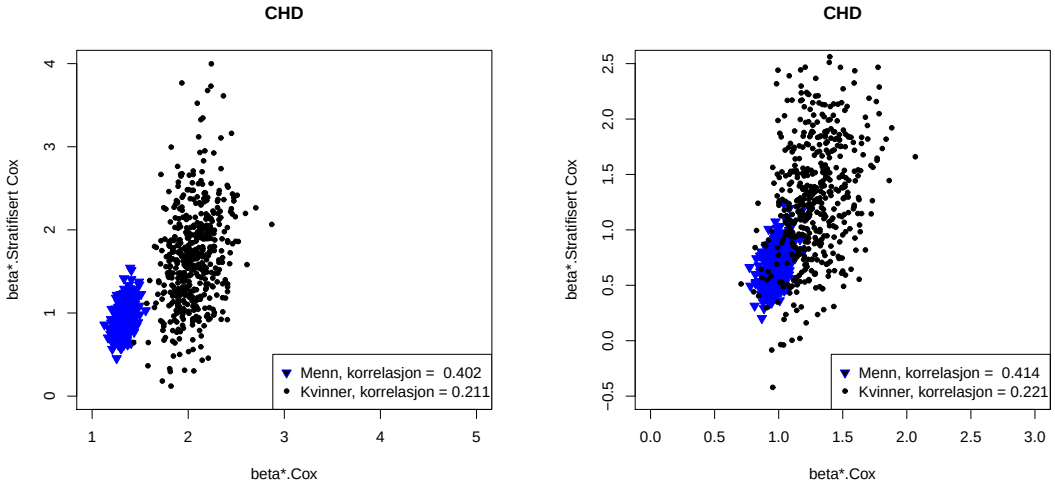


(a) (b)

Figur 5.2: (a) *12 års utdannelse: 500 bootstrap* (b) *12-16 års utdannelse: 500 bootstrap*

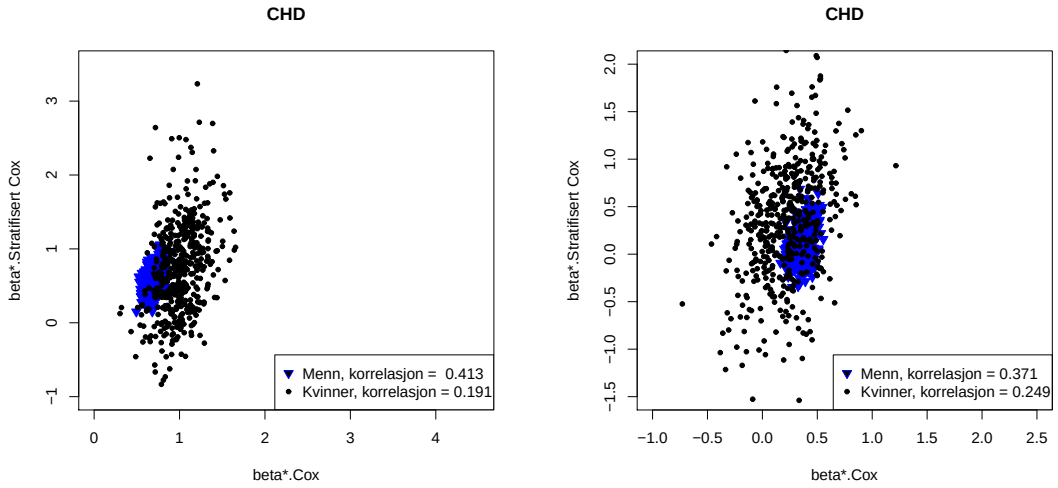
Tabell 5.3: *Sammenligning av Cox-modell med og uten stratifisering for CHD.*

		500 bootstrap-utvalg			
		Kvinner		Menn	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox					
	7-9 år	2.05	0.20	1.33	0.06
	10-11 år	1.29	0.20	0.98	0.06
	12 år	1.01	0.21	0.68	0.06
	12-16 år	0.23	0.26	0.36	0.07
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	1.81	1.72	0.94	0.16
	10-11 år	1.47	1.71	0.68	0.16
	12 år	0.88	1.72	0.55	0.16
	12-16 år	0.49	1.73	0.12	0.17
	>16 år	Ref.		Ref.	



(a) (b)

Figur 5.3: (a) *7-9 års utdannelse: 500 bootstrap* (b) *10-11 års utdannelse: 500 bootstrap*



(a) (b)

Figur 5.4: (a) *12 års utdannelse: 500 bootstrap* (b) *12-16 års utdannelse: 500 bootstrap*

5.3 Persentil-konfidensintervaller for totaldødelighet og utvalgte dødsårsaker

En måte å se om det er en signifikant forskjell mellom resultatene fra Cox regresjon og stratifisert Cox regresjon er å bruke persentilmetoden. Denne metoden er beskrevet i Kapittel 3.6. I Tabell 5.4 har vi funnet persentil-konfidensintervaller for forskjellene for 500 bootstrap-utvalg for totaldødeligheten og for CHD, alkoholrelaterte årsaker og lungekreft. I Tabell E.9 i Tillegg E.5 er de tilsvarende resultatene for 100 bootstrap-utvalg. Metoden er beskrevet i Seksjon 3.6.2. Hvis konfidensintervallet inneholder verdien 0, er det ikke noen signifikant forskjell på nivå 5% når vi tester forskjellen mellom metodene. Dersom det er tilfellet, kan det tyde på at metodene er like og familieavhengigheten ikke er viktig. I Tabell 5.4 inneholder alle intervallene verdien 0 for både totaldødelighet og de tre dødsårsakene for kvinner. Det er litt annerledes for menn. For totaldødeligheten dekker ikke konfidensintervallene 0 for de to første utdannelsesnivåene. For CHD gjelder det grunnskole, mens for alkoholrelaterte årsaker gjelder det for både grunnskole og høyskole. Blant dødsårsakene som er listet i tabellen, er det kun dødsårsaken lungekreft som ikke gir forkastning for alle utdannelsesnivåene når vi tester om det er forskjell mellom metodene.

5.4 Søskenflokker

I forrige seksjon har vi brukt bootstrap-utvalg for å teste om det er forskjell mellom Cox regresjonsmodell med og uten stratifisering på søskenflokk for hele datasettet. I denne seksjonen gjør vi det samme, men her tar vi ikke med ”enebarn”. I de søskenstratifiserte analysene vil derfor ikke ”enebarn” bidra. Det er interessant å se om resultatene endres

Tabell 5.4: *Konfidensintervaller for forskjellen mellom parameterne.*

		500 bootstrap-utvalg	
Dødsårsak	**Utdannelse**	**Kvinner**	**Menn**
Total-dødelighet			
	7-9 år	(-0.11,0.24)	(0.06,0.31)
	10-11 år	(-0.24,0.08)	(0.01 0.23)
	12 år	(-0.20,0.15)	(-0.02,0.19)
	12-16 år	(-0.21,0.13)	(-0.12,0.10)
	>16 år	Ref.	Ref.
CHD			
	7-9 år	(-1.19,1.41)	(0.09,0.66)
	10-11 år	(-1.42,0.92)	(-0.01,0.56)
	12 år	(-1.31,1.27)	(-0.17,0.38)
	12-16 år	(-1.31,1.27)	(-0.06,0.52)
	>16 år	Ref.	Ref.
Alkoholrelaterte årsaker			
	7-9 år	(-1.34,2.40)	(0.08,1.09)
	10-11 år	(-0.92,2.53)	(-0.02,0.98)
	12 år	(-1.16,2.30)	(0.08,1.17)
	12-16 år	(-1.64,1.60)	(-0.06,1.13)
	>16 år	Ref.	Ref.
Lungekreft			
	7-9 år	(-0.42,0.97)	(-0.50,0.62)
	10-11 år	(-0.81,0.58)	(-0.75,0.37)
	12 år	(-0.60,0.88)	(-0.64,0.41)
	12-16 år	(-0.66,0.65)	(-1.13,0.08)
	>16 år	Ref.	Ref.

ved å utelate "enebarn". Tabellene E.10, E.11 og E.12 i Tillegg E.6 viser resultatene for henholdsvis totaldødelighet, lungekreft og alkoholrelaterte årsaker. Fra disse tabellene ser vi blant annet at usikkerheten er større for den stratifiserte metoden. I tillegg har vi også en tabell som viser persentil-konfidensintervaller for forskjellen mellom disse metodene. Dette er vist i Tabell 5.5. Som nevnt i forrige seksjon, er det ikke en signifikant forskjell mellom de to metodene dersom intervallet inneholder verdien 0. For kvinner tyder resultatene fra Tabell 5.5 på at det ikke er noen forskjell mellom Cox regresjonsmodell med og uten stratifisering på søskenflokk, men ut fra tabellene i Tillegg E.6 ser det ut som det er forskjell for alkoholrelaterte årsaker. Her er standardfeilen for stratifisert Cox regresjon stor i forhold til Cox regresjon. Forskjellen er ikke like stor for lungekreft og totaldødelighet.

Tabell 5.5: *Konfidensintervaller for forskjellen mellom parameterne (søsken).*

		500 bootstrap-utvalg	
Dødsårsak	**Utdannelse**	**Kvinner**	**Menn**
Total dødelighet			
	7-9 år	(-0.12,0.19)	(0.06,0.27)
	10-11 år	(-0.27,0.02)	(-0.01,0.20)
	12 år	(-0.23,0.09)	(-0.03,0.17)
	12-16 år	(-0.22,0.09)	(-0.15,0.08)
	>16 år	Ref.	Ref.
Lungekreft			
	7-9 år	(-0.47,1.00)	(-0.39,0.76)
	10-11 år	(-0.74,0.60)	(-0.66,0.46)
	12 år	(-0.70,0.76)	(-0.55,0.50)
	12-16 år	(-0.68,0.67)	(-1.06,0.07)
	>16 år	Ref.	Ref.
Alkoholrelaterte årsaker			
	7-9 år	(-0.71,2.34)	(-0.03,1.01)
	10-11 år	(-0.71,2.30)	(-0.19,0.86)
	12 år	(-0.58,2.38)	(-0.06,0.98)
	12-16 år	(-1.03,1.63)	(-0.14,0.95)
	>16 år	Ref.	Ref.

Når det gjelder menn, er det kun det første persentilintervallet for totaldødelighet som ikke inneholder 0. Også her er det alkoholrelaterte årsaker som gir størst avvik mellom resultatene. Gjennomsnittet av de estimerte β-verdiene for Cox regresjon ble 1.68, 1.25, 0.70 og 0.31 for henholdsvis grunnskole, yrkesskole, gymnas/artium og høyskole. For den stratifiserte analysen ble verdiene lik 1.18, 0.86, 0.21 og -0.10. Dødsrisikoen blir mindre når vi stratifiserer på søskenflokk.

Kapittel 6

Resultater for tvillinger

I dette kapittelet har vi benyttet de samme metodene som i Kapittel 4 (Cox regresjon og frailty), men for tvillinger er det ikke foretatt egne analyser for kvinner og menn. Alkoholrelaterte årsaker er også ekskludert da dette var den dødsårsaken som hadde færrest observasjoner. Få observasjoner gir mer usikre estimater. Av de totalt 786 tvillingene som døde var det 22 menn og 6 kvinner som døde av alkoholrelaterte årsaker.

6.1 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering

Tabell 6.1 gir en oversikt over de estimerte hazard ratioene fra en tilpasset Cox regresjon, samt 95% konfidensintervaller for hazard ratioene. Vi har stratifisert på søskenflokk, slik som i forrige kapittel. For totaldødeligheten ser vi at dødsrisikoen minker med utdannelsesnivå og risikoen for død er dermed noe større for tvillinger med kun grunnskoleutdannelse. Konfidensintervallene for HR er bredere når vi stratifiserer på søskenflokk. Som tidligere nevnt, kommer det av at det kun er to personer i hvert stratum. I den stratifiserte analysen er det kun det første 95% konfidensintervallet som ikke inneholder verdien 1, og som dermed er signifikant. P-verdien for gruppen "7-9 år" ble 4%, som gir signifikans på nivå 5%. For gruppene "10-11 år", "12 år" og "12-16 år" ble de henholdsvis 13%, 66%, 20%. For *Cluster*, var det kun p-verdien for høyskole som ikke ble tilnærmet 0 når vi tester om hazard ratioen er lik 1.

For de øvrige dødsårsakene i Tabell 6.1 ble usikkerheten større. Konfidensintervallene for HR er en del bredere, og dette gjelder spesielt den stratifiserte analysen for ulykker. I det forrige kapittelet tok vi med hele datasettet i analysene, mens i Kapittel 6 er det 15 796 tvillinger som er tatt med i analysene. Som nevnt i Kapittel 2, svarer flerfødsler til 1.7% av hele datasettet. Som tidligere skrevet, blir estimatene mer sikre jo større datasettet er. Den største differansen mellom de estimerte hazard ratioene for høyeste og laveste utdannelsesnivå er for *cluster* for dødsårsakene CVD og CHD. I disse tilfellene ble dødsrisikoen omtrent seks ganger høyere for tvillinger med inntil niårig utdannelse i forhold til personer med universitetsutdannelse.

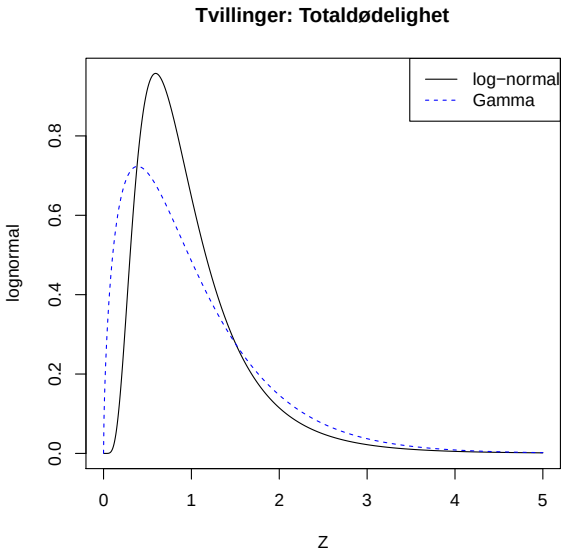
Tabell 6.1: *Cox regresjon med og uten stratifisering.*

		Strata		**Cluster**	
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**
Total-dødelighet					
	7-9 år	1.79	1.03-3.10	2.44	1.82-3.28
	10-11 år	1.62	0.96-2.73	1.75	1.30-2.35
	12 år	1.49	0.89-2.47	1.75	1.28-2.38
	12-16 år	1.14	0.65-1.98	1.14	0.79-1.65
	>16 år	Ref.		Ref.	
Lungekreft					
	7-9 år	1.28	0.19-8.68	5.30	1.64-1.72
	10-11 år	1.50	0.24-9.46	3.16	0.95-10.49
	12 år	1.11	0.14-8.68	1.94	0.53-7.01
	12-16 år	1.78	0.17-18.68	2.63	0.70-9.84
	>16 år	Ref.		Ref.	
CVD					
	7-9 år	3.72	0.79-17.40	6.04	2.17-16.79
	10-11 år	5.72	1.26-25.95	4.17	1.50-11.58
	12 år	4.30	1.04-17.70	5.73	2.05-16.04
	12-16 år	1.28	0.28-5.82	2.23	0.69-7.26
	>16 år	Ref.		Ref.	
CHD					
	7-9 år	2.17	0.32-14.96	6.07	1.86-19.82
	10-11 år	2.103	0.32-13.62	3.55	1.08-11.64
	12 år	2.38	0.41-13.72	4.59	1.38-15.19
	12-16 år	1.03	0.13-8.19	1.66	0.40-6.96
	>16 år	Ref.		Ref.	
Ulykker					
	7-9 år	5.32	0.46-61.24	3.31	1.14-9.57
	10-11 år	3.04	0.29-31.83	2.35	0.82-6.77
	12 år	3.07	0.28-33.63	2.21	0.74-6.60
	12-16 år	2.85	0.23-35.39	1.55	0.44-5.47
	>16 år	Ref.		Ref.	

6.2 Resultater fra gamma og log-normal frailty

For tvillinger er det større forskjell mellom metodene i Tabell 6.1 enn frailty-metodene gjengitt i Tabell 6.2. I tillegg er konfidensintervallene ikke like brede som de gjengitt i Tabell 6.1. Som vi så i Kapittel 4, gir gamma- og log-normal frailty tilnærmet like hazard ratioer. Som vi ser fra tabellene i dette kapittelet, gjelder gir gamma- og log-normal frailty tilnærmet like for tvillinger også.

For dødsårsakene lungekreft, CVD og CHD er det størst forskjell mellom dødsrisiko for de med lav utdannelse sammenlignet med de som har høyskole- og universitetsutdannelse. For både CVD og CHD er dødsrisikoen omtrent seks ganger høyere for gruppen med grunnskoleutdannelse sammenlignet med gruppen med universitetsutdannelse. Når utdannelsesnivået øker, minker risikoen for død. Også her har vi funnet θ for gamma- og log-normal frailty, som er *variance of random effect*. Fra Tabell 6.3 kan man se at det er lungekreft som har størst verdi av alle dødsårsakene. Dette gjelder begge fordelingene. Variansen ble 5.53 og 1.38 for henholdsvis gamma og log-normal frailty. Variansen viser at det er forskjell mellom de to fordelingene selv om resultatene fra Tabell 6.2 ikke tyder på det. For alle dødsårsakene er θ_G-verdiene er høyere enn θ_{LN}-verdien. I tillegg er tetthetsplottene i figurene 6.1, 6.2 og 6.3 ulike for gamma- og log-normalfordeling der θ_G- og θ_{LN}-verdiene fra `R` er satt inn i formlene for fordelingene.



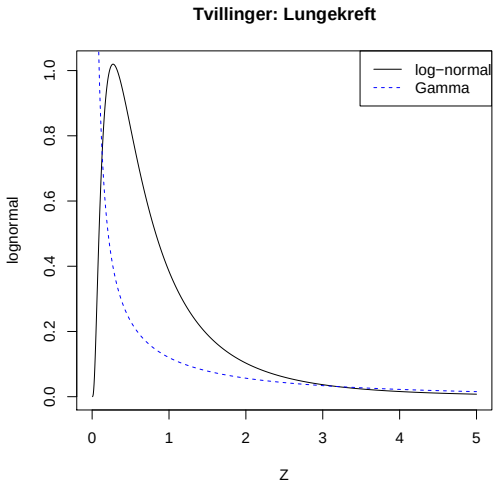
Figur 6.1: $\theta_G = 0.61$, $1\theta_{LN} = 0.42$

Tabell 6.2: *Gamma og Invers Gaussisk frailty.*

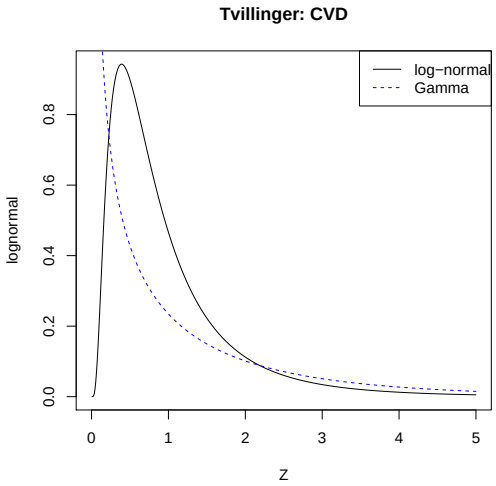
		Gamma		Log-normal	
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**
Total-dødelighet					
	7-9 år	2.48	1.83-3.36	2.46	1.82-3.32
	10-11 år	1.77	1.31-2.40	1.76	1.30-2.38
	12 år	1.76	1.28-2.41	1.75	1.28-2.40
	12-16 år	1.15	0.79-1.67	1.15	0.79-1.66
	>16 år	Ref.		Ref.	
Lungekreft					
	7-9 år	5.45	1.65-18.09	5.27	1.61-17.30
	10-11 år	3.25	0.97-10.91	3.17	0.95-10.53
	12 år	1.94	0.53-7.14	1.92	0.53-7.01
	12-16 år	2.65	0.70-10.15	2.63	0.69-9.94
	>16 år	Ref.		Ref.	
CVD					
	7-9 år	6.08	2.18-16.98	6.01	2.16-16.71
	10-11 år	4.25	1.52-11.90	4.19	1.50-11.70
	12 år	5.84	2.08-16.42	5.76	2.06-16.13
	12-16 år	2.23	0.68-7.28	2.22	0.68-7.24
	>16 år	Ref.		Ref.	
CHD					
	7-9 år	6.08	1.87-19.8	6.06	1.86-19.17
	10-11 år	3.56	1.08-11.73	3.55	1.08-11.69
	12 år	4.60	1.38-15.32	4.59	1.38-15.26
	12-16 år	1.66	0.40-6.97	1.66	0.40-6.96
	>16 år	Ref.		Ref.	
Ulykker					
	7-9 år	3.35	1.15-9.79	3.31	1.14-9.60
	10-11 år	2.37	0.82-6.87	2.35	0.81-6.78
	12 år	2.22	0.74-6.70	2.21	0.74-6.62
	12-16 år	1.55	0.44-5-54	1.55	0.44-5.49
	>16 år	Ref.		Ref.	

Tabell 6.3: *Variance of random effect for gamma- og log-normal frailty.*

	Gamma	Log-normal
Dødsårsak	$\boldsymbol{\theta_G}$	$\boldsymbol{\theta_{LN}}$
Totaldødelighet	0.61	0.42
Lungekreft	5.53	1.38
CVD	2.10	0.87
CHD	1.07	0.60
Ulykker	4.12	1.16

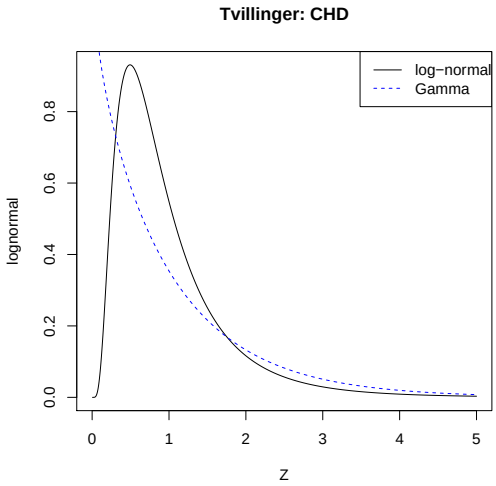


(a)

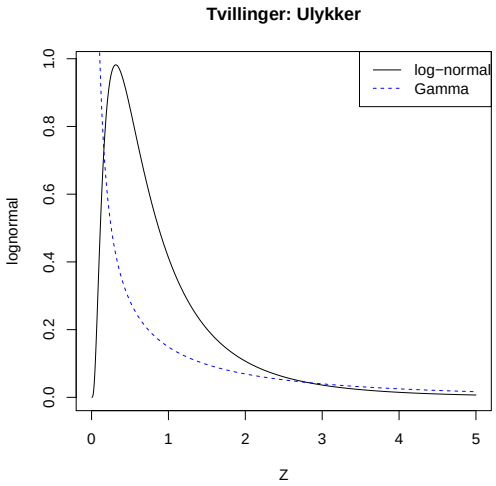


(b)

Figur 6.2: (a) $\theta_G = 5.53$, $\theta_{LN} = 1.38$ (b) $\theta_G = 2.10$, $\theta_{LN} = 0.87$



(a)



(b)

Figur 6.3: (a) $\theta_G = 1.07$, $\theta_{LN} = 0.60$ (b) $\theta_G = 4.12$, $\theta_{LN} = 1.16$

6.3 Resultater fra gamma og log-normal frailty for tvillinger med samme kjønn

For å kunne sammenligne resultater for tvillinger med resultatene i Kapittel 4, har vi tatt med egne frailty-analyser for tvillinger splittet på kjønn. På grunn av få observasjoner, har vi kun studert totaldødelighet og CVD, som var den dødsårsaken flest tvillinger døde av. Av 15 769 tvillinger, var det 113 menn og 36 kvinner som døde på grunn av CVD. Tabell 4.4 i Kapittel 4 gjengir de samme resultatene for hele datasettet. Igjen er det slik at det er ikke er stor forskjell mellom gamma- og log-normal frailty. En sammenligning av frailty-resultatene viser at de estimerte hazard-ratioene for kvinner er høyere når vi bruker hele datasettet enn for tvillinger. Det er motsatt for menn. Der er hazard-ratioene for tvillinger høyere, som vil si at dødsrisikoen er høyere for menn.

Når det gjelder *variance of random effect* i Tabell 6.3 og Tabell 4.3, ble θ-verdiene høyere for menn her sammenlignet med Tabell 4.3 for både totaldødelighet og CVD. Dette gjaldt både gamma- og log-normal frailty. Jo høyere verdi av θ, jo mer avhengighet er det innad i gruppene. I dette tilfellet kan det tyde på at det er større avhengighet blant tvillinger (menn) enn søsken generelt. For eksempel ble θ_G lik 3.22 for CVD for tvillinger som var menn, mens det tilsvarende tallet for hele data ble 1.29, og indikerer større avhengighet blant tvillinger enn søsken. Denne variansen ble litt annerledes for kvinner som er tvillinger. Her ser vi at θ_G og θ_{LN} for CVD er tilnærmet 0. For totaldødeligheten ble θ_G og θ_{LN} høyere i Tabell 6.3 enn Tabell 4.3. Resultatene for CVD ble usikre, det kan igjen skyldes lite datagrunnlag for tvillinger og få dødsfall blant kvinner som er tvillinger.

Tabell 6.4: *Variance of random effect for gamma- og log-normal frailty tvillinger med samme kjønn.*

	θ_G: Gamma		θ_{LN}: Log-normal	
Dødsårsak	**Kvinner**	**Menn**	**Kvinner**	**Menn**
Totaldødelighet	1.280	0.820	0.633	0.486
CVD	$5 * 10^{-5}$	3.220	0.003	1.070

I Tabell 6.5 vises hazard ratioer og 95% konfidensintervaller for totaldødelighet. Resultatene for CVD er ikke tatt med her på grunn av betydelig usikkerhet i estimatene. Standardavvikene for estimatene ble store i forhold til de resultatene der hele datasettet ble brukt. Sammenlignet med Tabell 4.4 i Kapittel 4.4 er hazard ratioene for kvinner lavere her, mens konfidensintervallene er bredere for både gamma- og log-normal. For menn ble både estimatene og standardavvikene høyere.

Tabell 6.5: *Gamma- og log-normal frailty for tvillinger med samme kjønn.*

		Gamma				**Log-normal**			
		Kvinner		**Menn**		**Kvinner**		**Menn**	
Dødsårsak	**Utd.**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**
Total-dødelighet									
	7-9 år	1.48	0.92-2.38	3.68	2.47-5.50	1.46	0.91-2.33	3.61	2.43-5.37
	10-11 år	1.22	0.77-1.93	2.61	1.74-3.92	1.20	0.76-1.90	2.58	1.72-3.86
	12 år	0.65	0.36-1.19	2.41	1.61-3.60	0.66	0.37-1.19	2.39	1.60-3.56
	12-16 år	0.87	0.49-1.55	1.41	0.85-2.33	0.87	0.50-1.54	1.40	0.85-2.31
	>16 år	Ref.		Ref.		Ref.		Ref.	

Kapittel 7

Oppsummering og videre arbeid

7.1 Konklusjon

I denne oppgaven har vi sett på sammenhengen mellom antall år med utdannelse og årsaksspesifikk dødelighet innen søskenflokker Vi har sammenlignet flere metoder som tar hensyn til avhengighet mellom søsken på ulike måter. De metoden som er brukt er Cox regresjonsmodell og frailty-modeller. Dette har vi gjort for totaldødelighet og ulike dødsårsaker. Hovedsakelig har vi egne analyser for kvinner og menn, men de stedene dette ikke er blitt gjort er grunnet få observasjoner som ville gitt usikre resultater. Som nevnt i Kapittel 2, bestod datasettet av totalt 934 548 personer. Med et datasett av dette omfanget blir resultatene mer pålitelige.

I Kapittel 4 har vi brukt hele datasettet til å undersøke om familieavhengigheten er viktig. Der viste resultatene ingen markant forskjell mellom Cox regresjon med og uten sandwich-estimator for 95% konfidensintervaller for hazard ratioen. Dette gjaldt både menn og kvinner. Som beskrevet i Kapittel 3, gir disse metodene like koeffisienter, men ulike konfidensintervaller på grunn av varians-estimatoren. Fra Tabell 4.1 så vi blant annet at den største forskjellen mellom de høyt- og lavtutdannede gruppene, der dødsrisikoen ble kraftig redusert med utdannelsesnivå, gjaldt for CHD og alkoholrelaterte årsaker for kvinner. Sistnevnte dødsårsak gjaldt også for menn i tillegg til lungekreft. Videre så vi på analyser for Cox regresjon med stratifisering på søskenflokk. Ved å stratifisere vil hver av de 292 300 søskenflokkene ha en baseline hazard og man antar at det er avhengighet innad i hver søskenflokk og søskenflokkene er uavhengige av hverandre. Disse resultatene viste at for kvinner var det størst forskjell for den gruppen med kun grunnskole for dødsårsakene lungekreft, CHD og alkoholrelaterte årsaker. Dødsrisikoen ble redusert sammenlignet med samme modell uten stratifisering på søskenflokk. For eksempel var hazard ratioen for alkoholrelaterte årsaker lik 6.9 og 2.6 for kvinner for henholdsvis. Cox regresjonsmodell med og uten stratifisering på søskenflokk. Resultatene viser samme trend for menn. Resultatene viste også at standardavvikene ble større når vi tok hensyn til familieavhengigheten, som da førte til bredere konfidensintervaller. Dette skyldes flere forhold, som er nevnt i Kapittel 4.2.

Vi har også sett på andre metoder som tar hensyn til familieavhengigheten på en annen måte og har brukt gamma- og log-normal frailty. Hazard ratioene og konfidensintervallene ble omtrent like for disse to metodene, men plott for tetthetene og *variance of random effect* i Kapittel 4.3 viste stor forskjell mellom gamma- og log-normal frailty.

For å se om det var noen forskjell mellom resultatene for de ulike metodene, sammenlignet vi de fire metodene: Cox regresjon med og uten stratifisering og gamma- og log-normal frailty, og konsentrerte oss om totaldødelighet, CHD, alkoholrelaterte årsaker og lungekreft. Grunnen var at det var disse dødsårsakene som hadde de høyeste dødsrisikoene i de innledende analysene. Her ble resultatene for de fire metodene tilnærmet like, men standardavvikene ble høyere når vi stratifiserte på søskenflokk i forhold til de andre metodene. For CHD ble $\hat{\beta}$-verdiene fra stratifisert Cox regresjon lavere.

I Kapittel 4.6 tilpasset vi en regresjonsmodell for "enebarn" og søsken. For sistnevnte ble familieavhengighet mellom søsken hensyntatt. Dette er lignende analyser som i artikkelen *Education and adult cause specific mortality-examining the impact of family factors shared by 871 367 Norwegian siblings*[1]. Antall "enebarn" var 212 743 og i analysene ble de regnet som uavhengige. Konfidensintervallene for "enebarn" ble bredere. Vi testet om det var en signifikant forskjell mellom de to gruppene for de ulike dødsårsakene. På nivå 5% ble CHD og alkoholrelaterte årsaker signifikante. Ved å sammenligne med resultatene fra artikkelen, er det noen avvik, men som tidligere nevnt, er det færre personer i det datasettet som brukes her sammenlignet med det som ble brukt i artikkelen. For totaldødelighet ble resultatene tilnærmet like, men for de andre årsakene ble det noe forskjell.

For å teste de resultatene vi har fått, har vi brukt bootstrapping. I Kapittel 5 har vi funnet gjennomsnitt og standardfeil for bootstrap-estimatene. Hovedsakelig er det estimater for 500 bootstrap-utvalg, men vi testet også resultatene for 100 bootstrap-utvalg. I Tabell 5.4 og Tabell 5.5 tester vi forskjellen mellom Cox regresjonsmodell med og uten stratifisering på søskenflokk med persentil-metoden. Dette er gjort for datasettet med og uten "enebarn". Også her valgte vi å fokusere på totaldødelighet, CHD, alkoholrelaterte årsaker og lungekreft på grunn av at de estimerte hazard ratioene var høyere sammenlignet med de andre dødsårsakene. I det første tilfellet, der vi tok med "enebarn" i analysene, viste persentilintervallene at det ikke er forskjell mellom de to metodene for kvinner for de dødsårsakene som er nevnt. For menn er det gruppen med inntil 9-årig skole for dødsårsakene CHD og alkoholrelaterte årsaker som gir forkastning. I den analysen der vi kun tok med søskenflokker ble konklusjonen lik for kvinner. For menn ble det ikke signifikant forskjell for alkoholrelaterte årsaker. Vi sammenlignet disse resultatene med de fra Kapittel 4 der Cox regresjons ble brukt. Bootstrap-gjennomsnittene og estimatene fra Cox regresjon ble tilnærmet like.

For tvillinger er resultatene mer usikre. De utgjorde 1.7% av datasettet. Vi tok ikke med alkoholrelaterte årsaker fordi det var få tvillinger som døde av dette. I tillegg har vi analyser for begge kjønn sammen. I Kapittel 6.3 har vi analyser separat for kvinner og menn for CVD og totaldødelighet. Grunnen til at vi tok CVD var at det var fordi det var den årsaken som forårsaket flest dødsfall blant de tvillingene i datasettet. Vi har gjort dette for å sammenligne resultatene for tvillinger med resultatene for hele datasettet. Som for

resten av datasettet, ble resultatene fra gamma- og log-normal frailty like og tetthetsplottene ulike.

Totalt sett ser det ikke ut som søskenavhengigheten er viktig siden vi ikke fant noen systematiske forskjeller. De fire forskjellige metodene ga tilnærmet de samme resultatene. For alle dødsårsakene er det slik at dødsrisikoen er størst for de gruppene med lavest utdannelse, og risikoen avtar når utdannelsesnivået øker. Siden vi gjennom hele oppgaven har brukt universitet som referansegruppe, er det de med høyskoleutdannelse som har lavest risiko sammenlignet med denne gruppen. Det er også slik at menn har høyere risiko enn kvinner.

7.2 Videre arbeid

I denne oppgaven har vi brukt bootstrapping for å teste resultatene for søsken. Dette har vi ikke gjort for "enebarn" eller tvillinger. Når det gjelder analysene for søsken, kan man justere for flere variable, som for eksempel antall i husholdet. Innledningsvis i Kapittel 4 har vi justert for fødselsår. Dette ga ikke en signifikant endring av resultatene, men det er uvisst om noen av de andre variablene vil føre til signifikant forskjellige resultater.

Temaet for denne oppgaven er dagsaktuelt og er mye diskutert i media. Livsforsikringsselskapene har stor fokus på dødelighet og må ta høyde for at levealderen i Norge øker. I mars 2013 bestemte Finanstilsynet at det skal innføres et nytt dødelighetsgrunnlag[1]. Sikkerhetsmarginer er lagt til for å hensynta sosioøkonomiske forhold.

Resultatene i denne oppgaven er vist til noen av mine kolleger i Mercer og de har allerede henvist til resultatet i denne oppgaven i sine prosjekter og har kommentert at de tror at flere oppdrag innen dette fagområdet vil komme.

[1] www.finanstilsynet.no

Tillegg A

Datasettet

A.1 Antall personer i datasettet delt inn etter fødselsår

Tabell A.1: *Personene i datasettet delt inn etter fødselsår.*

	Antall		Total
Fødselsår	**Menn**	**Kvinner**	**andel**
1940-1944	87 401	68 51	16.7%
1945-1949	128 077	120 511	26.6%
1950-1954	129 710	123 618	27.1%
1955-1959	141 075	135 641	29.6%

Tabell A.2: *Antall personer som døde delt inn etter fødselsår.*

	Antall		Total
Fødselsår	**Menn**	**Kvinner**	**andel**
1940-1944	10 456	5 040	31%
1945-1949	10 155	6 382	33%
1950-1954	6 545	3 949	21%
1955-1959	4 783	2 783	15%
Totalt	31 938	18 117	100%

A.2 Tabeller med antall døde og rater for totaldødelighet, CHD, alkoholrelaterte årsaker og lungekreft

Tabell A.3: *Rater per 1 000 separat for kvinner og menn for totaldødelighet.*

Alders-	Rate per 100 personår	
gruppe	Menn	Kvinner
30-34 år	0.132	0.064
35-39 år	0.433	0.195
40-44 år	1.143	0.633
45-49 år	2.275	1.365
50-54 år	4.122	2.455
55-59 år	6.147	3.640
60-64 år	9.253	5.052
65-69 år	17.441	7.682

Tabell A.4: *Antall døde pga CHD og rate per 1 000 separat for kvinner og menn.*

	Menn		Kvinner	
Alders-		Rate per		Rate per
gruppe	Antall	1000 personår	Antall	1000 personår
30-34 år	5	0.006	0	0.000
35-39 år	81	0.036	14	0.006
40-44 år	328	0.147	60	0.027
45-49 år	864	0.388	132	0.059
50-54 år	1 438	0.767	267	0.142
55-59 år	1 333	1.095	267	0.219
60-64 år	935	1.514	203	0.329
65-69 år	355	2.690	77	0.583

Tabell A.5: *Antall døde pga alkoholrelaterte årsaker og rater per 1 000 separat for kvinner og menn.*

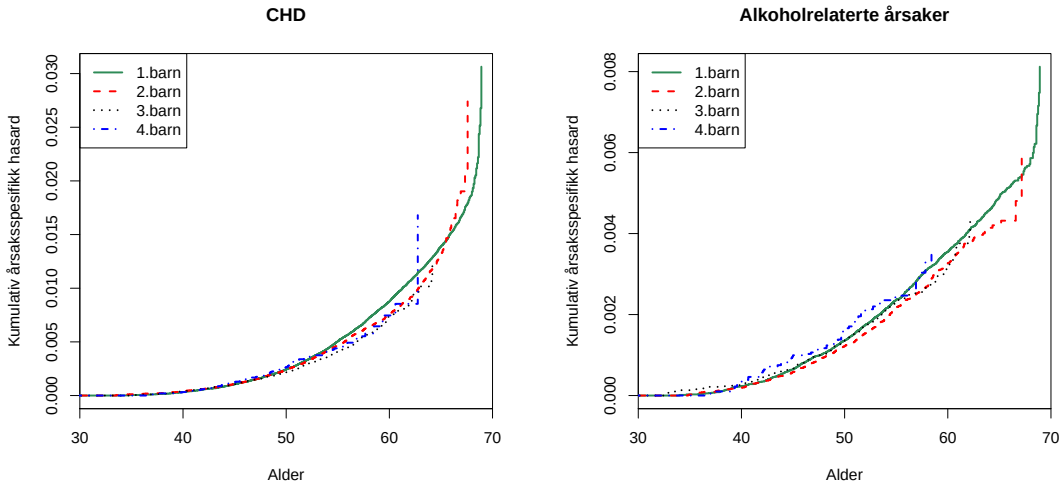
	Menn		Kvinner	
Alder	**Antall**	**Rate per 1000 personår**	**Antall**	**Rate per 1000 personår**
30-34 år	3	0.003	1	0.001
35-39 år	46	0.021	15	0.007
40-44 år	183	0.082	43	0.019
45-49 år	423	0.190	95	0.043
50-54 år	533	0.284	168	0.090
55-59 år	435	0.357	114	0.094
60-64 år	228	0.369	69	0.112
65-69 år	71	0.538	13	0.098
Totalt	1922		518	

Tabell A.6: *Antall døde pga lungekreft og rater per 1 000 separat for kvinner og menn.*

	Menn		Kvinner	
Alder	**Antall**	**Rate per 1000 personår**	**Antall**	**Rate per 1000 personår**
30-34 år	0	0.000	0	0.000
35-39 år	6	0.003	8	0.004
40-44 år	79	0.035	82	0.037
45-49 år	260	0.117	218	0.098
50-54 år	619	0.330	435	0.232
55-59 år	770	0.633	543	0.446
60-64 år	666	1.078	403	0.652
65-69 år	296	2.243	130	0.985
Totalt	2 696		1 819	

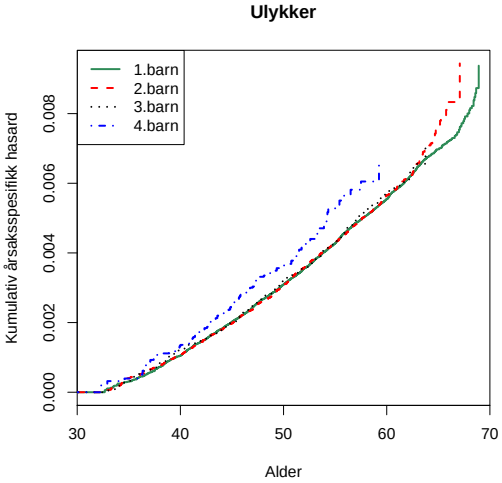
Tillegg B

Nelson-Aalen plott for CHD, alkoholrelaterte årsaker og ulykker



(a) (b)

Figur B.1: *Nelson-Aalen*

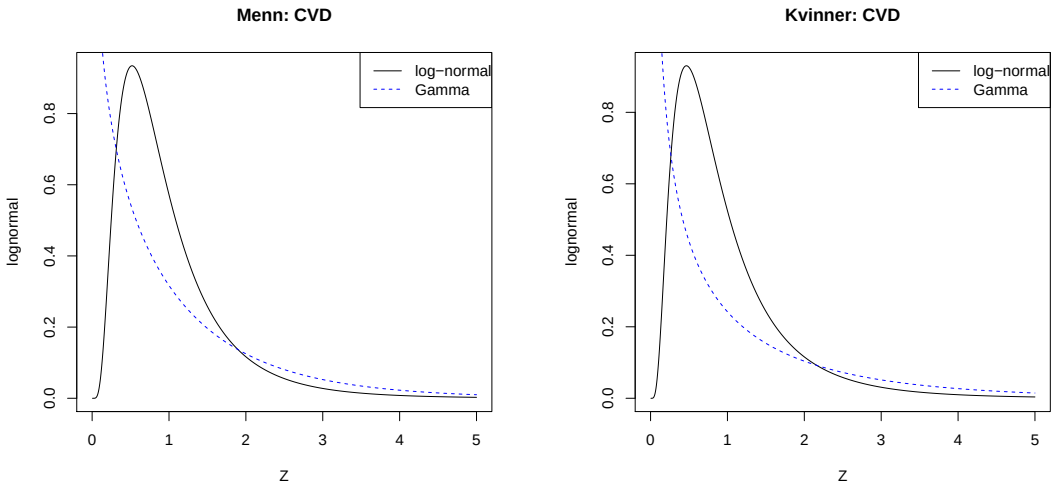


Figur B.2: *Nelson-Aalen*

Tillegg C

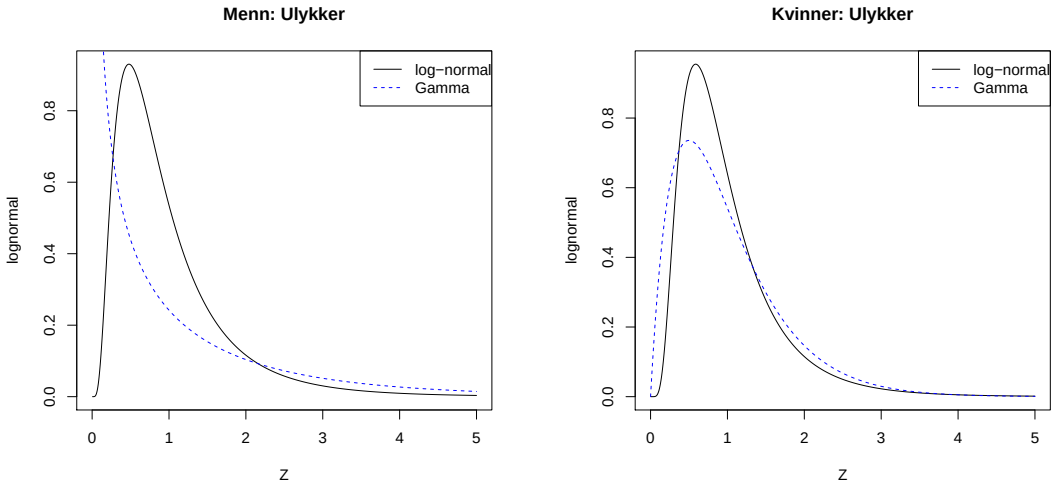
Resultater fra frailty-analysene

C.1 Tetthetsplott for gamma- og log-normal frailty



(a) (b)

Figur C.1: (a) $\theta_G = 1.29$, $\theta_{LN} = 0.55$ (b) $\theta_G = 2.00$, $\theta_{LN} = 0.63$



(a) (b)

Figur C.2: (a) $\theta_G = 2.00$, $\theta_{LN} = 0.63$ (b) $\theta_G = 0.50$, $\theta_{LN} = 0.43$

C.2 Resultater fra gamma- log-normal frailty for hele datasettet

Tabell C.1: *Gamma frailty inkludert "enebarn".*

		Kvinner			Menn		
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**
Total-dødelighet							
	7-9 år	2.21	0.03	2.07-2.35	2.90	0.02	2.77-3.03
	10-11 år	1.51	0.03	1.42-1.60	2.24	0.02	2.14-2.34
	12 år	1.38	0.04	1.28-1.48	1.67	0.02	1.59-1.74
	12-16 år	1.14	0.05	1.06-1.23	1.32	0.03	1.25-1.40
	>16 år	Ref.			Ref.		
Lungekreft							
	7-9 år	5.07	0.13	3.95-6.49	5.57	0.09	4.62-6.72
	10-11 år	2.56	0.13	2.00-3.29	3.87	0.10	3.20-4.69
	12 år	2.11	0.14	1.60-2.79	2.98	0.10	2.46-3.61
	12-16 år	1.14	0.16	0.84-1.56	2.18	0.11	1.75-2.72
	>16 år	Ref.			Ref.		
CVD							
	7-9 år	5.31	0.12	4.17-6.77	3.60	0.05	3.27-3.96
	10-11 år	3.02	0.12	2.37-3.84	2.62	0.05	2.37-2.88
	12 år	2.49	0.14	1.91-3.24	1.92	0.05	1.74-2.12
	12-16 år	1.48	0.15	1.11-1.98	1.40	0.06	1.24-1.58
	>16 år	Ref.			Ref.		
CHD							
	7-9 år	7.70	0.20	5.25-11.30	3.85	0.06	3.44-4.32
	10-11 år	3.58	0.20	2.43-5.26	2.69	0.06	2.39-3.02
	12 år	2.71	0.21	1.78-4.12	1.99	0.06	1.77-2.24
	12-16 år	1.24	0.25	0.77-2.01	1.44	0.07	1.25-1.67
	>16 år	Ref.			Ref.		
Alkoholrelaterte årsaker							
	7-9 år	6.92	0.26	4.16-11.50	6.03	0.11	4.85-7.49
	10-11 år	3.32	0.26	2.00-5.53	3.99	0.11	3.20-4.98
	12 år	2.37	0.29	1.35-4.16	2.41	0.12	1.92-3.02
	12-16 år	1.29	0.32	0.68-2.43	1.58	0.12	1.20-2.08
	>16 år	Ref.			Ref.		
Ulykker							
	7-9 år	2.27	0.14	1.72- 2.99	3.38	0.08	2.90-3.94
	10-11 år	1.35	0.14	1.03-1.78	2.47	0.08	2.12-2.89
	12 år	1.30	0.16	0.96-1.77	1.77	0.08	1.51-2.06
	12-16 år	0.95	0.18	0.67-1.33	1.17	0.10	0.96-1.42
	>16 år	Ref.			Ref.		

Tabell C.2: *Log-normal frailty inkludert "enebarn".*

		Kvinner			Menn		
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**
Total-dødelighet							
	7-9 år	2.20	0.03	2.07-2.34	2.89	0.02	2.77-3.02
	10-11 år	1.51	0.03	1.42-1.60	2.24	0.02	2.14-2.34
	12 år	1.37	0.04	1.28-1.47	1.66	0.02	1.59-1.74
	12-16 år	1.14	0.04	1.06-1.23	1.32	0.03	1.25-1.40
	>16 år	Ref.			Ref.		
Lungekreft							
	7-9 år	5.01	0.13	3.91-6.41	5.53	0.10	4.59-6.66
	10-11 år	2.55	0.13	1.99-3.27	3.85	0.10	3.19-4.66
	12 år	2.11	0.14	1.60-2.78	2.97	0.10	2.46-3.59
	12-16 år	1.14	0.16	0.84-1.56	2.18	0.11	1.75-2.72
	>16 år	Ref.			Ref.		
CVD							
	7-9 år	5.27	0.12	4.14-6.71	3.55	0.05	3.23-3.91
	10-11 år	3.00	0.12	2.36-3.82	2.60	0.05	2.36-2.86
	12 år	2.48	0.14	1.90-3.23	1.91	0.05	1.73-2.11
	12-16 år	1.48	0.15	1.11-1.98	1.40	0.06	1.24-1.58
	>16 år	Ref.			Ref.		
CHD							
	7-9 år	7.60	0.20	5.18-11.14	3.80	0.06	3.39-4.26
	10-11 år	3.55	0.20	2.42-5.22	2.67	0.06	2.37-2.99
	12 år	2.70	0.21	1.77-4.10	1.98	0.06	1.76-2.22
	12-16 år	1.24	0.25	0.77-2.01	1.44	0.07	1.25-1.66
	>16 år	Ref.			Ref.		
Alkoholrelaterte årsaker							
	7-9 år	6.91	0.26	4.16-11.50	5.96	0.11	4.80-7.40
	10-11 år	3.32	0.26	2.00-5.53	3.96	0.11	3.18-4.94
	12 år	2.37	0.29	1.35-4.16	2.40	0.12	1.92-3.00
	12-16 år	1.29	0.32	0.68-2.43	1.58	0.14	1.20-2.08
	>16 år	Ref.			Ref.		
Ulykker							
	7-9 år	2.27	0.14	1.72-2.99	3.36	0.08	2.89-3.91
	10-11 år	1.35	0.14	1.03-1.78	2.46	0.08	2.11-2.87
	12 år	1.30	0.16	0.96-1.77	1.76	0.08	1.51-2.06
	12-16 år	0.95	0.18	0.67-1.33	1.16	0.10	0.96-1.42
	>16 år	Ref.			Ref.		

C.3 Resultater fra gamma frailty uten "enebarn"

Tabell C.3: *Gamma frailty uten "enebarn".*

		Kvinner			Menn		
Dødsårsak	**Utdannelse**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**	$\widehat{\mathbf{HR}}$	**se**$(\hat{\beta})$	**95% k.i**
Total dødelighet							
	7-9 år	2.15	0.04	2.00-2.31	2.85	0.03	2.71-3.00
	10-11 år	1.45	0.04	1.35-1.55	2.18	0.03	2.07- 2.29
	12 år	1.31	0.04	1.21-1.42	1.63	0.03	1.55-1.72
	12-16 år	1.13	0.04	1.03-1.22	1.30	0.03	1.22-1.39
	>16 år	Ref.			Ref.		
Lungekreft							
	7-9 år	4.95	0.14	3.73-6.57	6.11	0.11	4.88-7.64
	10-11 år	2.67	0.14	2.01-3.54	4.06	0.12	3.23-5.11
	12 år	2.06	0.16	1.50- 2.83	3.19	0.12	2.54-4.01
	12-16 år	1.19	0.18	0.83-1.69	2.28	0.14	1.75-2.97
	>16 år	Ref.			Ref.		
CVD							
	7-9 år	5.27	0.14	3.98-6.97	3.45	0.06	3.09-3.85
	10-11 år	2.87	0.14	2.17-3.80	3.45	0.06	2.22-2.78
	12 år	2.61	0.16	1.93-3.55	1.86	0.06	1.66-2.08
	12-16 år	1.42	0.17	1.01-1.99	1.32	0.07	1.15-1.52
	>16 år	Ref.			Ref.		
CHD							
	7-9 år	7.69	0.23	4.94-11.97	3.60	0.07	3.16-4.11
	10-11 år	3.36	0.23	2.15-5.24	2.55	0.07	2.23-2.92
	12 år	2.71	0.25	1.66-4.41	1.93	0.07	1.68-2.20
	12-16 år	1.23	0.29	0.70-2.15	1.32	0.09	1.12-1.56
	>16 år	Ref.			Ref.		
Alkoholrelaterte årsaker							
	7-9 år	6.70	0.30	3.73-12.06	5.39	0.12	4.23-6.87
	10-11 år	2.67	0.30	1.48-4.85	3.49	0.13	2.73-4.47
	12 år	2.22	0.34	1.15-4.28	2.00	0.13	1.55-2.58
	12-16 år	1.37	0.37	0.66-2.82	1.37	0.09	1.00-1.88
	>16 år	Ref.			Ref.		
Ulykker							
	7-9 år	2.08	0.16	1.53- 2.83	3.20	0.09	2.69-3.80
	10-11 år	1.25	0.16	0.92-1.70	2.41	0.09	2.02- 2.86
	12 år	1.16	0.18	0.82-1.66	1.72	0.09	1.44-2.05
	12-16 år	0.96	0.19	0.66-1.40	1.13	0.11	0.90-1.42
	>16 år	Ref.			Ref.		

C.4 Sammenligning av Cox regresjonsmodeller og frailty-modeller for alkoholrelaterte årsaker

Tabell C.4: *Sammenligning av metoder for alkoholrelaterte årsaker.*

		Kvinner		**Menn**	
Metode	**Utdannelse**	$\hat{\beta}$	**se**($\hat{\beta}$)	$\hat{\beta}$	**se**($\hat{\beta}$)
Cox					
	7-9 år	1.93	0.26	1.78	0.11
	10-11 år	1.20	0.26	1.38	0.11
	12 år	0.86	0.29	0.88	0.12
	12-16 år	0.25	0.32	0.46	0.14
	>16 år	Ref.		Ref.	
Stratifisert Cox (Stratifisert på søskenflokk)					
	7-9 år	0.97	0.72	1.19	0.28
	10-11 år	0.11	0.67	0.87	0.27
	12 år	0.04	0.66	0.23	0.27
	12-16 år	0.04	0.71	-0.08	0.31
	>16 år	Ref.		Ref.	
Frailty: Gamma					
	7-9 år	1.93	0.26	1.80	0.11
	10-11 år	1.20	0.26	1.38	0.11
	12 år	0.86	0.29	0.88	0.12
	12-16 år	0.25	0.32	0.46	0.14
	>16 år	Ref.		Ref.	
Frailty: Log-normal					
	7-9 år	1.93	0.26	1.79	0.11
	10-11 år	1.20	0.26	1.38	0.11
	12 år	0.86	0.29	0.89	0.12
	12-16 år	0.25	0.32	0.46	0.14
	>16 år	Ref.		Ref.	

C.5 Sammenligning av Cox regresjonsmodeller og frailty-modeller for lungekreft

Tabell C.5: *Sammenligning av metoder for lungekreft.*

		Kvinner		Menn	
Metode	**Utdannelse**	$\hat{\beta}$	**se**($\hat{\beta}$)	$\hat{\beta}$	**se**($\hat{\beta}$)
Cox					
	7-9 år	1.61	0.13	1.71	0.09
	10-11 år	0.94	0.13	1.35	0.10
	12 år	0.75	0.14	1.09	0.10
	12-16 år	0.13	0.16	0.78	0.11
	>16 år	Ref.		Ref.	
Stratifisert Cox (Stratifisert på søskenflokk)					
	7-9 år	1.23	0.37	1.58	0.29
	10-11 år	0.95	0.36	1.45	0.28
	12 år	0.58	0.37	1.16	0.27
	12-16 år	0.13	0.37	1.24	0.30
	>16 år	Ref.		Ref.	
Frailty: Gamma					
	7-9 år	1.62	0.13	1.72	0.10
	10-11 år	0.94	0.13	1.35	0.10
	12 år	0.75	0.14	1.09	0.10
	12-16 år	0.14	0.16	0.78	0.11
	>16 år	Ref.		Ref.	
Frailty: Log-normal					
	7-9 år	1.61	0.13	1.71	0.10
	10-11 år	0.94	0.13	1.35	0.10
	12 år	0.75	0.14	1.09	0.10
	12-16 år	0.14	0.16	0.78	0.11
	>16 år	Ref.		Ref.	

Tillegg D

Resultater for "enebarn" og søsken

D.1 Resultater fra Cox regresjon separat for "enebarn" og søsken

Tabell D.1: *Cox regresjon uten stratifisering for "enebarn" og søsken. For søsken er det tatt hensyn til familieavhengighet ved å bruke sandwich.estimatoren.*

		"Enebarn"				Søsken			
Døds-		**Kvinner**		**Menn**		**Kvinner**		**Menn**	
årsak	**Utd.**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**	$\widehat{\mathbf{HR}}$	**95% k.i**
CVD									
	7-9 år	5.39	3.32- 8.73	3.97	3.28-4.82	5.23	3.95-6.92	3.41	3.06-3.80
	10-11 år	3.37	2.09-5.44	2.95	2.43-3.59	2.86	2.16-3.78	2.47	2.21-2.76
	12 år	2.11	1.24-3.59	2.09	1.72-2.55	2.60	1.92-3.53	1.85	1.65-2.07
	12-16 år	1.64	0.93-2.89	1.65	1.31-2.00	1.42	1.01-1.99	1.32	1.15-1.51
	>16 år	Ref.		Ref.		Ref.		Ref.	
CHD									
	7-9 år	7.64	3.56-16.42	4.60	3.63-5.83	7.60	4.88-11.83	3.56	3.13-4.05
	10-11 år	4.13	1.93-8.86	3.09	2.43-3.93	3.34	2.14-5.20	2.53	2.22-2.89
	12 år	2.67	1.16-6.12	2.18	1.70-2.78	2.69	1.66-4.38	1.92	1.68-2.19
	12-16 år	1.28	0.50-3.31	1.84	1.38-2.44	1.23	0.70-2.14	1.32	1.12-1.55
	>16 år	Ref.		Ref.		Ref.		Ref.	
Ulykker									
	7-9 år	3.24	1.74-6.04	4.04	2.90-5.62	2.08	1.53-2.83	3.19	2.69-3.78
	10-11 år	1.84	1.00-3.39	2.71	1.94-3.78	1.25	0.92-1.70	2.40	2.02-2.85
	12 år	1.86	0.96-3.60	1.93	1.38-2.70	1.16	0.82-1.66	1.72	1.44-2.05
	12-16 år	0.89	0.40-1.99	1.29	0.85-1.96	0.96	0.66-1.40	1.13	0.90-1.42
	>16 år	Ref.		Ref.		Ref.		Ref.	

Tabell D.2: *Cox regresjon uten stratifisering for "enebarn" og søsken. For søsken er det tatt hensyn til familieavhengighet ved å bruke sandwich-estimatoren.*

		Kvinner		Menn	
Dødsårsak	**Utdannelse**	**z**	**p-verdi**	**z**	**p-verdi**
Totaldødelighet					
	7-9 år	3.519	0.000	4.893	0.000
	10-11 år	3.542	0.000	5.348	0.000
	12 år	2.928	0.003	2.516	0.011
	12-16 år	0.919	0.358	1.403	0.161
	>16 år	Ref.			
Lungekreft					
	7-9 år	1.612	0.107	-8.318	0.000
	10-11 år	-1.401	0.161	-3.097	0.002
	12 år	0.579	0.563	-3.258	0.001
	12-16 år	-0.473	0.636	-1.327	0.185
	>16 år	Ref.			
CVD					
	7-9 år	0.545	0.586	4.967	0.000
	10-11 år	1.797	0.072	4.279	0.000
	12 år	-1.585	0.113	2.093	0.036
	12-16 år	0.670	0.503	2.430	0.015
	>16 år	Ref.			
CHD					
	7-9 år	0.100	0.920	7.555	0.000
	10-11 år	1.766	0.077	3.998	0.000
	12 år	-0.058	0.954	1.822	0.068
	12-16 år	0.103	0.918	3.091	0.002
	>16 år	Ref.			
Alkohol-relaterte årsaker	7-9 år	1.608	0.108	12.152	0.000
	10-11 år	4.236	0.000	9.211	0.000
	12 år	0.787	0.431	7.506	0.000
	12-16 år	-0.412	0.680	3.312	0.001
	>16 år	Ref.			
Ulykker					
	7-9 år	3.290	0.001	4.484	0.000
	10-11 år	1.689	0.091	1.632	0.103
	12 år	1.825	0.068	1.074	0.283
	12-16 år	-0.147	0.883	0.671	0.502
	>16 år	Ref.			

Tillegg E

Bootstrapping

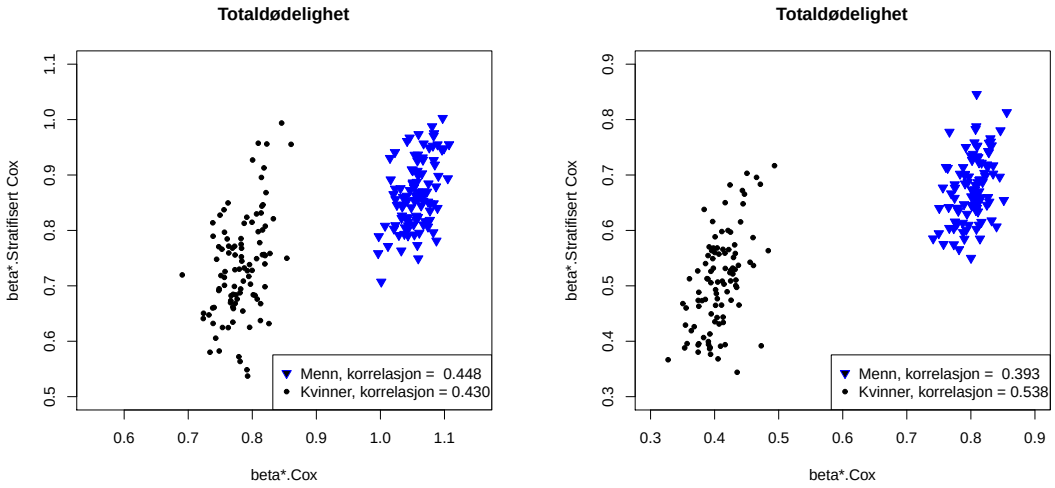
E.1 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering for totaldødelighet

Tabell E.1: *Sammenligning av Cox modell med og uten stratifisering for totaldødelighet.*

		100 bootstrap-utvalg			
		Kvinner		**Menn**	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox					
	7-9 år	0.78	0.03	1.06	0.02
	10-11 år	0.41	0.03	0.80	0.02
	12 år	0.32	0.04	0.51	0.02
	12-16 år	0.13	0.04	0.28	0.03
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	0.74	0.09	0.86	0.06
	10-11 år	0.51	0.09	0.68	0.06
	12 år	0.36	0.09	0.42	0.06
	12-16 år	0.17	0.09	0.29	0.06
	>16 år	Ref.		Ref.	

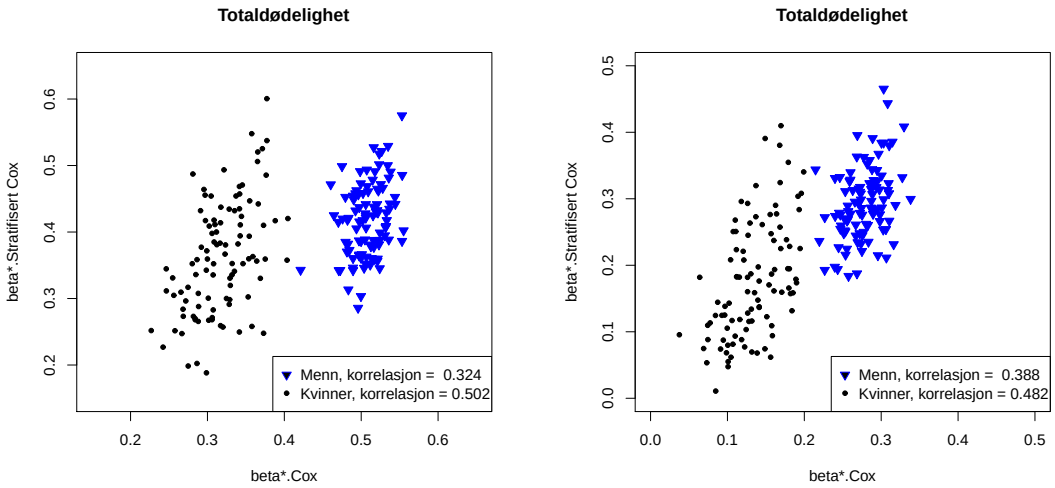
Tabell E.2: *Variansen for totaldødelighet.*

	100 bootstrap-utvalg	
Utdannelse	**Kvinner**	**Menn**
7-9 år	0.007	0.003
10-11 år	0.006	0.003
12 år	0.005	0.003
12-16 år	0.007	0.003
>16 år	Ref.	Ref.



(a) (b)

Figur E.1: (a) *7-9 års utdannelse: 100 bootstrap* (b) *10-11 års utdannelse: 100 bootstrap*



(a) (b)

Figur E.2: (a) *12 års utdannelse: 100 bootstrap* (b) *12-16 års utdannelse: 100 bootstrap*

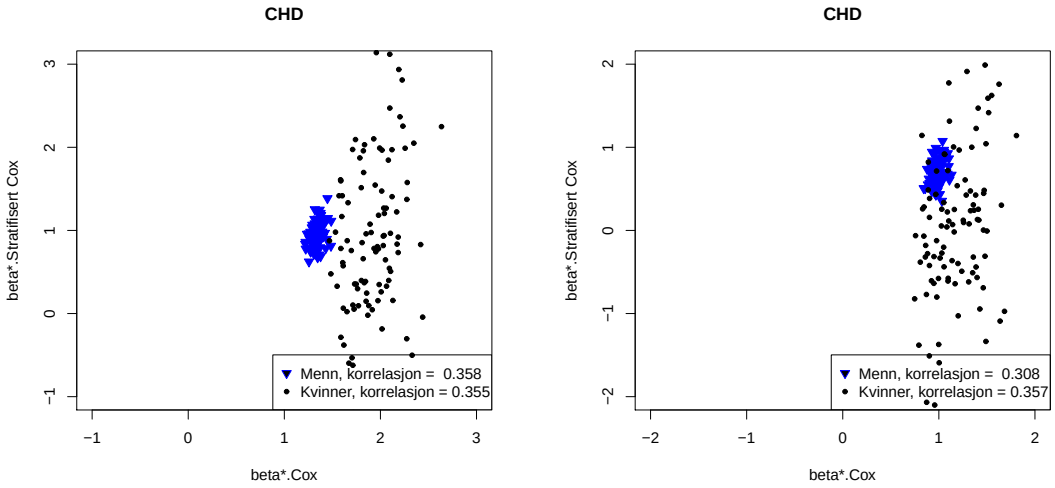
E.2 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering for CHD

Tabell E.3: *Sammenligning av Cox modell med og uten stratifisering for CHD.*

		100 bootstrap-utvalg			
		Kvinner		**Menn**	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox					
	7-9 år	1.93	0.25	1.34	0.06
	10-11 år	1.18	0.25	0.98	0.06
	12 år	0.84	0.30	0.69	0.06
	12-16 år	0.24	0.32	0.37	0.08
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	1.11	1.94	0.95	0.15
	10-11 år	0.25	1.93	0.69	0.15
	12 år	0.06	1.79	0.55	0.14
	12-16 år	0.003	1.83	0.15	0.17
	>16 år	Ref.		Ref.	

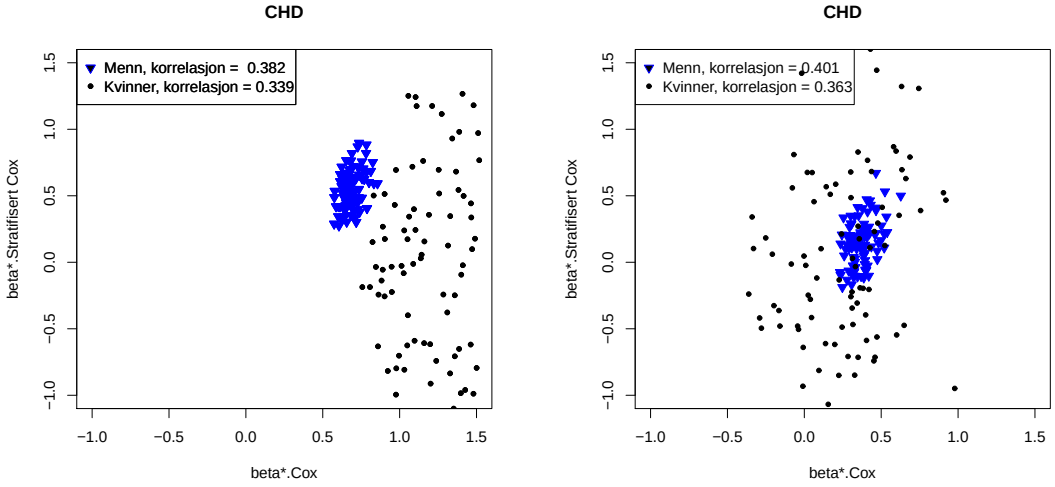
Tabell E.4: *Variansen for CHD.*

	100 bootstrap-utvalg	
Utdannelse	**Kvinner**	**Menn**
7-9 år	3.482	0.019
10-11 år	3.424	0.020
12 år	2.963	0.017
12-16 år	3.015	0.017
>16 år	Ref.	Ref.



(a) (b)

Figur E.3: (a) *7-9 års utdannelse: 100 bootstrap* (b) *10-11 års utdannelse: 100 bootstrap*



(a) (b)

Figur E.4: (a) *12 års utdannelse: 100 bootstrap* (b) *12-16 års utdannelse: 100 bootstrap*

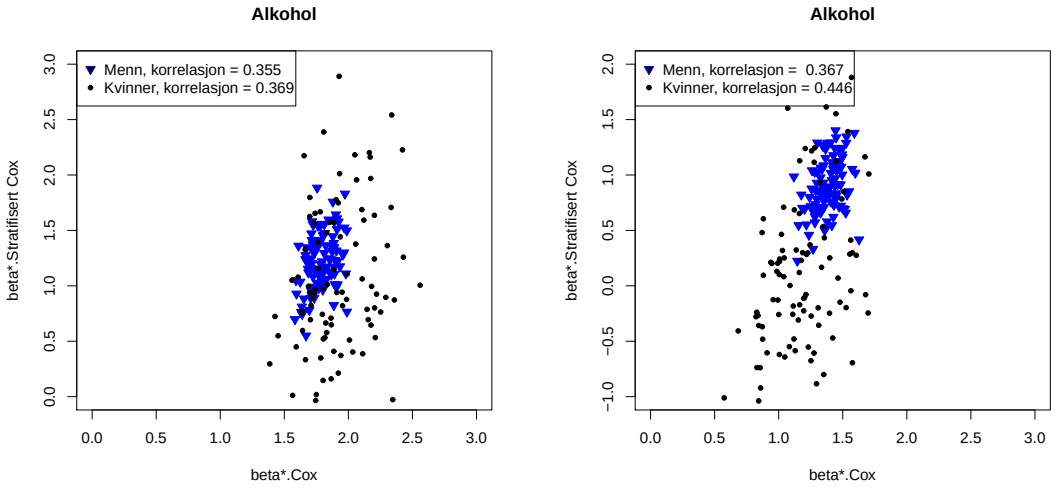
E.3 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering for alkoholrelaterte årsaker.

Tabell E.5: *Sammenligning av Cox modell med og uten stratifisering for alkoholrelaterte årsaker.*

		100 bootstrap-utvalg				**500 bootstrap-utvalg**			
		Kvinner		**Menn**		**Kvinner**		**Menn**	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\beta}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox									
	7-9 år	1.90	0.26	1.79	0.10	1.98	0.30	1.78	0.11
	10-11 år	1.17	0.26	1.38	0.10	1.25	0.30	1.37	0.11
	12 år	0.80	0.26	0.89	0.10	0.91	0.32	0.88	0.11
	12-16 år	0.25	0.31	0.45	0.13	0.27	0.38	0.45	0.14
	>16 år	Ref.		Ref.		Ref.		Ref.	
Stratifisert Cox (søskenflokk)									
	7-9 år	0.92	0.75	1.22	0.26	1.23	2.10	1.17	0.29
	10-11 år	0.08	0.75	0.90	0.25	0.37	2.07	0.86	0.28
	12 år	-0.07	0.82	0.26	0.24	0.27	2.04	0.21	0.28
	12-16 år	0.12	0.78	-0.03	0.30	0.25	2.03	-0.11	0.32
	>16 år	Ref.		Ref.		Ref.		Ref.	

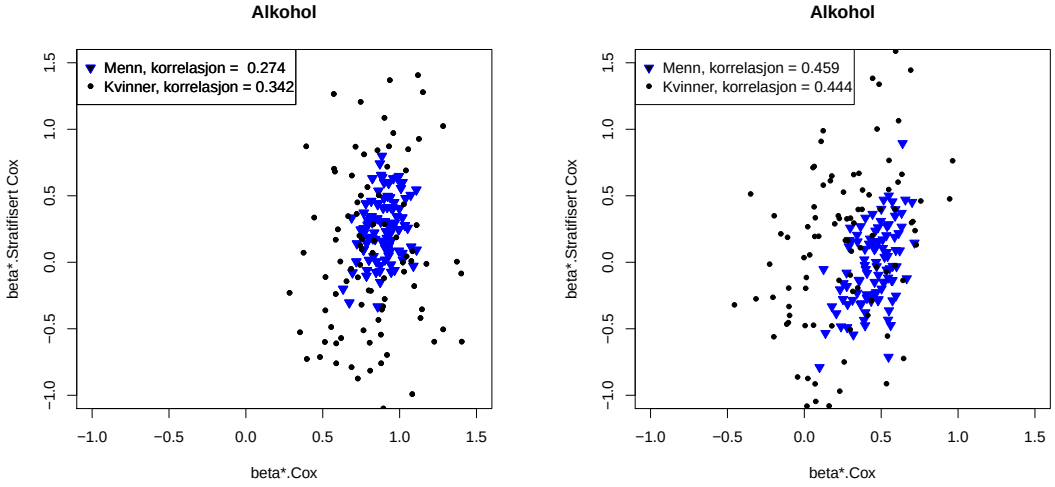
Tabell E.6: *Variansen for alkoholrelaterte årsaker.*

	100 bootstrap-utvalg		**500 bootstrap-utvalg**	
Utdannelse	**Kvinner**	**Menn**	**Kvinner**	**Menn**
7-9 år	0.486	0.057	4.097	0.073
10-11 år	0.459	0.054	3.990	0.066
12 år	0.0541	0.591	3.849	0.072
12-16 år	0.484	0.071	3.732	0.085
>16 år	Ref.	Ref.	Ref.	Ref.



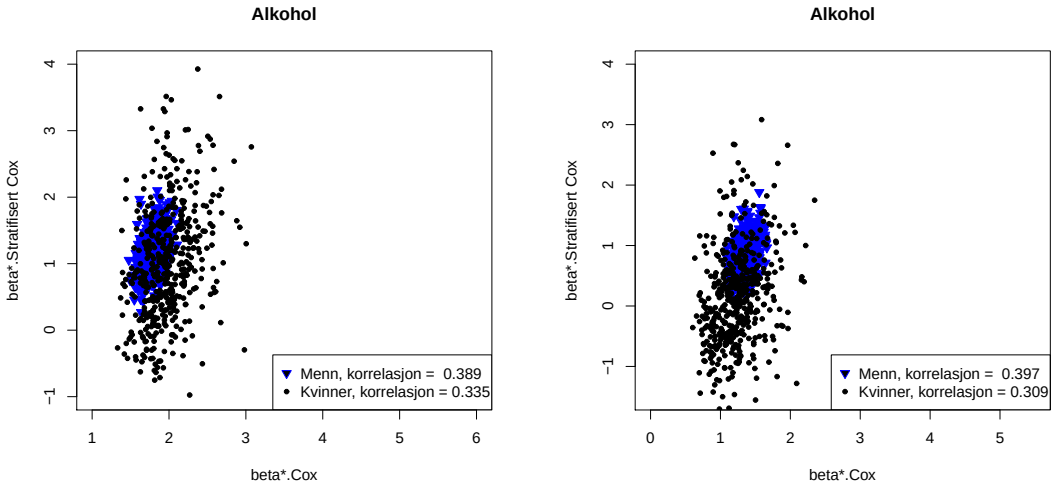
(a) (b)

Figur E.5: (a) *7-9 års utdannelse: 100 bootstrap* (b) *10-11 års utdannelse: 100 bootstrap*



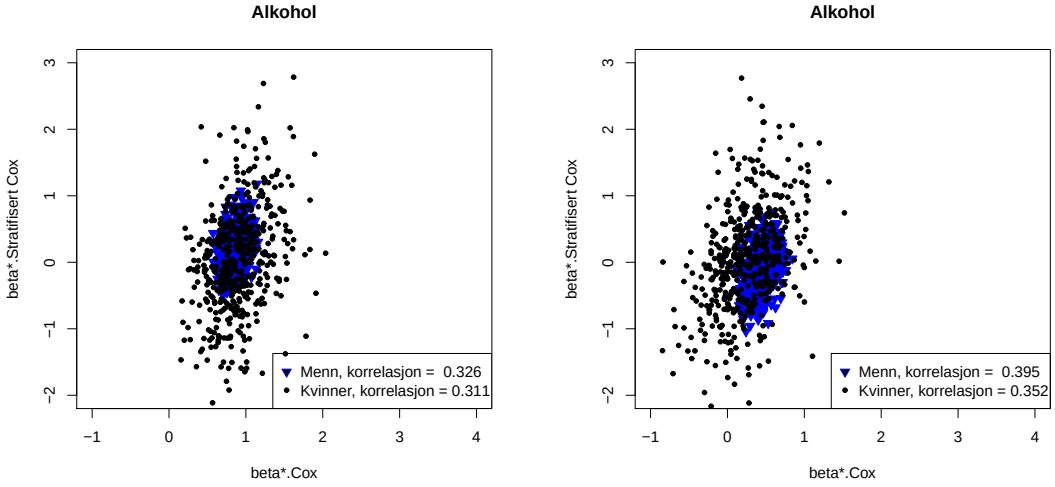
(a) (b)

Figur E.6: (a) *12 års utdannelse: 100 bootstrap* (b) *12-16 års utdannelse: 100 bootstrap*



(a) (b)

Figur E.7: (a) *7-9 års utdannelse: 500 bootstrap* (b) *10-11 års utdannelse: 500 bootstrap*



(a) (b)

Figur E.8: (a) *12 års utdannelse: 500 bootstrap* (b) *12-16 års utdannelse: 500 bootatrap*

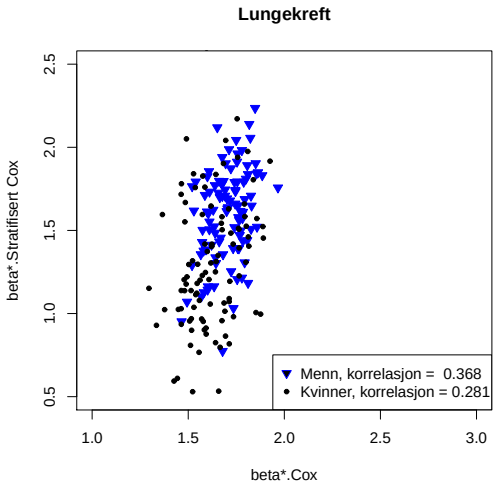
E.4 Resultater fra tilpasset Cox regresjonsmodell med og uten stratifisering for lungekreft

Tabell E.7: *Sammenligning av Cox modell med og uten stratifisering for lungekreft.*

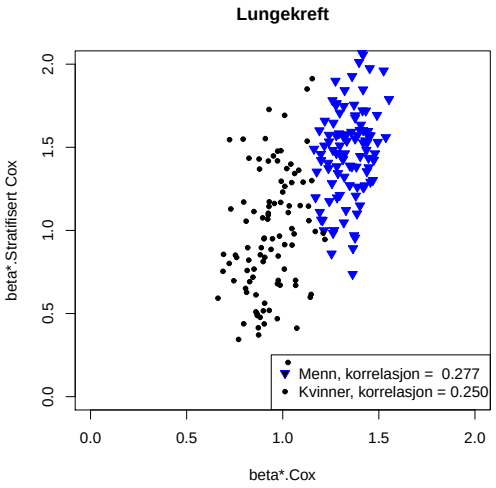
		100 bootstrap-utvalg				**500 bootstrap-utvalg**			
		Kvinner		**Menn**		**Kvinner**		**Menn**	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\beta}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox									
	7-9 år	1.62	0.13	1.70	0.10	1.62	0.13	1.71	0.10
	10-11 år	0.94	1.12	1.34	0.09	0.94	0.13	1.35	0.10
	Gym2	0.75	0.15	1.08	0.10	0.75	0.15	1.09	0.10
	12-16 år	0.15	0.15	0.76	0.11	0.13	0.17	0.78	0.11
	>16 år	Ref.		Ref.		Ref.		Ref.	
Stratifisert Cox (søskenflokk)									
	7-9 år	1.31	0.39	1.60	0.28	1.26	0.39	1.59	0.29
	10-11 år	0.99	0.39	1.46	0.28	0.96	0.37	1.47	0.28
	Gym2	0.64	0.43	1.17	0.27	0.58	0.40	1.17	0.27
	12-16 år	0.20	0.39	1.21	0.30	0.12	0.37	1.26	0.31
	>16 år	Ref.		Ref.		Ref.		Ref.	

Tabell E.8: *Variansen for lungekreft.*

	100 bootstrap-utvalg		**500 bootstrap-utvalg**	
Utdannelse	**Kvinner**	**Menn**	**Kvinner**	**Menn**
7-9 år	0.138	0.066	0.129	0.077
10-11 år	0.143	0.070	0.115	0.075
12 år	0.171	0.065	0.135	0.066
12-16 år	0.153	0.074	0.114	0.086
>16 år	Ref.	Ref.	Ref.	Ref.

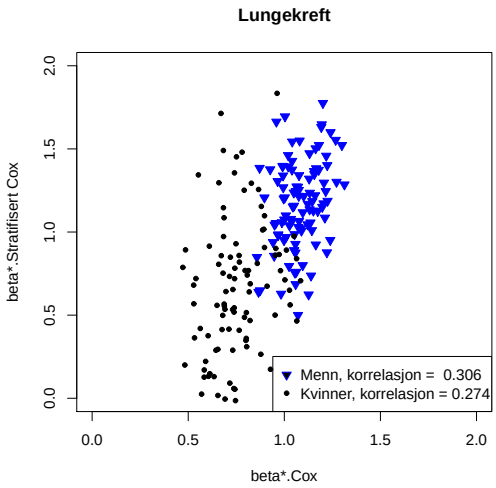


(a)

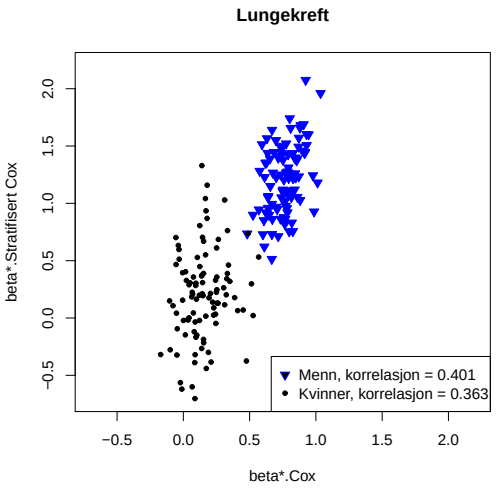


(b)

Figur E.9: (a) *7-9 års utdannelse: 100 bootstrap* (b) *10-11 års utdannelse: 100 bootstrap*

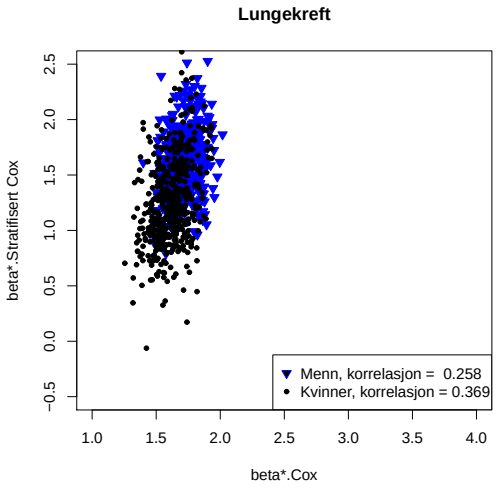


(a)

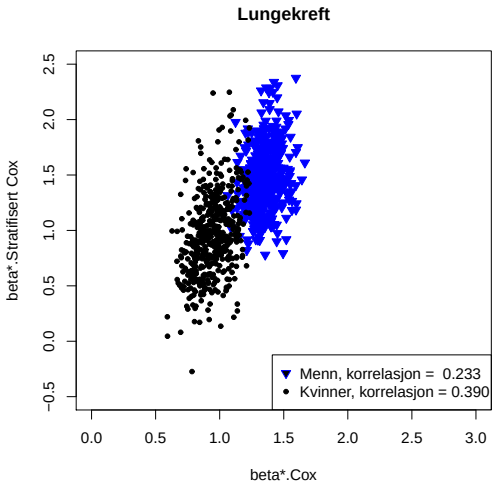


(b)

Figur E.10: (a) *12 års utdannelse: 100 bootstrap* (b) *12-16 års utdannelse: 100 bootstrap*

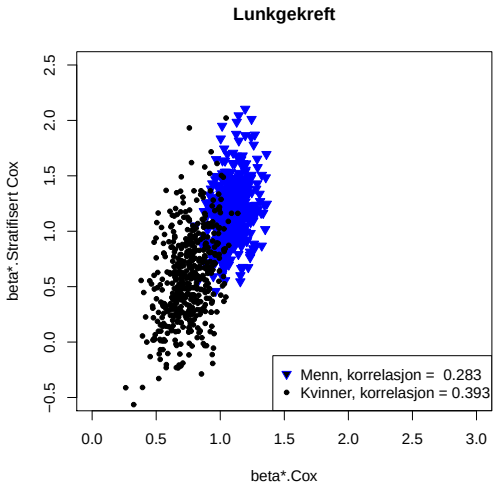


(a)

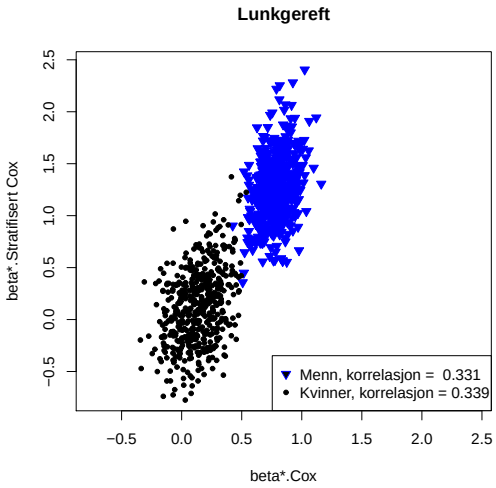


(b)

Figur E.11: (a) *7-9 års utdannelse: 500 bootstrap* (b) *10-11 års utdannelse: 500 bootstrap*



(a)



(b)

Figur E.12: (a) *12 års utdannelse: 500 bootstrap* (b) *12-16 års utdannelse: 500 bootstrap*

E.5 Persentil-konfidensintervaller for totaldødelighet og utvalgte dødsårsaker

Tabell E.9: *Konfidensintervaller for forskjellen mellom parameterne.*

		100 bootstrap-utvalg	
Dødsårsak	**Utdannelse**	**Kvinner**	**Menn**
Total-dødelighet			
	7-9 år	(-0.15,0.21)	(0.08,0.29)
	10-11 år	(-0.25,0.02)	(-0.01,0.22)
	12 år	(-0.21,0.09)	(-0.02,0.17)
	12-16 år	(-0.24,0.07)	(-0.13,0.08)
	>16 år	Ref.	Ref.
CHD			
	7-9 år	(-1.18,2.48)	(0.06,0.65)
	10-11 år	(-0.94,2.73)	(-0.02,0.52)
	12 år	(-0.89,2.45)	(-0.14,0.36)
	12-16 år	(-1.44,1.93)	(-0.08,0.47)
	>16 år	Ref.	Ref.
Alkoholrelaterte årsaker			
	7-9 år	(-0.58,2.15)	(0.12,0.92)
	10-11 år	(-0.53,2.18)	(0.04,0.87)
		(-0.69,2.28)	(0.13,1.02)
	12-16 år	(-1.45,1.45)	(0.04,0.91)
	>16 år	Ref.	Ref.
Lungekreft			
	7-9 år	(-0.56,0.88)	(-0.39,0.56)
	10-11 år	(-0.82,0.53)	(-0.64,0.40)
	12 år	(-0.87,0.76)	(-0.69,0.37)
	12-16 år	(-0.98,0.61)	(-0.97,0.04)
	>16 år	Ref.	Ref.

E.6 Resultater fra bootstrapping for totaldødelighet, alkoholrelaterte årsaker og lungekreft for søskenflokker

Tabell E.10: *Sammenligning av Cox modell med og uten stratifisering for søsken for totaldødelighet.*

		500 bootstrap-utvalg			
		Kvinner		**Menn**	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox					
	7-9 år	0.76	0.04	1.04	0.03
	10-11 år	0.37	0.03	0.77	0.03
	12 år	0.27	0.04	0.49	0.03
	12-16 år	0.12	0.04	0.26	0.03
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	0.72	0.09	0.87	0.06
	10-11 år	0.50	0.08	0.68	0.06
	12 år	0.35	0.09	0.42	0.06
	12-16 år	0.18	0.09	0.30	0.06
	>16 år	Ref.		Ref.	

Tabell E.11: *Sammenligning av Cox modell med og uten stratifisering for søsken for alkoholrelaterte årsaker.*

		500 bootstrap-utvalg			
		Kvinner		Menn	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox					
	7-9 år	1.92	0.31	1.68	0.12
	10-11 år	1.00	0.31	1.25	0.12
	12 år	0.80	0.34	0.70	0.13
	12-16	0.32	0.39	0.31	0.16
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	0.91	1.69	1.18	0.29
	10-11 år	0.02	1.68	0.86	0.30
	12 år	-0.09	1.69	0.21	0.30
	12-16 år	-0.06	1.63	-0.10	0.35
	>16 år	Ref.		Ref.	

Tabell E.12: *Sammenligning av Cox modell med og uten stratifisering for søsken for lungekreft.*

		500 bootstrap-utvalg			
		Kvinner		Menn	
Metode	**Utdannelse**	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$	$\bar{\hat{\beta}}^*$	$s_{\hat{\beta}}$
Cox					
	7-9 år	1.61	0.15	1.80	0.12
	10-11 år	0.99	0.15	1.39	0.12
	12 år	0.73	0.16	1.15	0.12
	12-16 år	0.18	0.19	0.82	0.14
	>16 år	Ref.		Ref.	
Stratifisert Cox (søskenflokk)					
	7-9 år	1.28	0.38	1.59	0.31
	10-11 år	0.99	0.37	1.46	0.30
	12 år	0.61	0.39	1.17	0.29
	12-16 år	0.15	0.40	1.26	0.33
	>16 år	Ref.		Ref.	

Bibliografi

[1] Øyvind Næss, Dominic A Hoff, Debbie Lawlor, Laust H Mortensen (2012). *Education and adult cause specific mortality-examining the impact of family factors shared by 871 367 Norwegian siblings*, International Journal of Epidemiology, 1-9.

[2] Hosmer D.W , Lemeshow S.(1998). *Applied Survival Analysis*. Canada. John Wiley & Sons, 12:28:33-34:317-319.

[3] John P.Klein, Melvin L.Moeschberger (2003) . *Survival Analysis: Techniques for Censored and Truncated data*. USA. Springer, 22:27:63:72:243-245:253-254:308:425:436.

[4] Odd O.Aalen, Ørnulf Borgan, Håkon K.Gjessing (2008). *Survival and Event History Analysis*. USA. Springer, 4:90-91:94:134-135:148:232-237:271:275-277:241-244.

[5] Duchateau L. , Janssen P. (2007). *The Frailty Model*. USA. Springer, 17-18:44:118-119:150-151.

[6] Terry M.Therneau, Patricia M. Granbsch (2000). *Modeling Survival Data: Extending the Cox Model*. USA. Springer, 172-173.

[7] Jay L. Devore, Kenneth N. Berk (2007). *Modern Mathematical Statistics with Applications*. USA. Springer, 404-409.

[8] Cox, D. R. (1972) *Regression models and life-tables (with discussion)*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 34, 187-220.